Analyse de sensibilité pour des modèles stochastiques à entrées dépendantes : application en énergétique du bâtiment

Mathilde Grandjacques

UNIVERSITÉ GRENOBLE ALPES

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

Spécialité : **Génie électrique**

Arrêté ministériel : 7 Aout 2006

Présentée par

Mathilde Grandjacques

Thèse dirigée par **Benoit Delinchant**
et codirigée par **Olivier Adrot**

préparée au sein **G2ELab : Grenoble Electrical Engineering Laboratory**
et de **l'école doctorale d'Electronique, Electrotechnique, Automatique et Traitement du Signal (EEATS)**

Analyse de sensibilité pour des modèles stochastiques à entrées dépendantes : application en énergétique du bâtiment

Thèse soutenue publiquement le **9 Novembre 2015**,
devant le jury composé de :

Madame, Clémentine Prieur
Professeur à l'Université de Grenoble, Présidente
Monsieur, Stéphane Clenet
Professeur ENSAM Lille, Rapporteur
Monsieur, Bertrand Iooss
Ingénieur de recherche EDF Chatou, Rapporteur
Monsieur, Benoit Delinchant
Maître de conférence à l'Université de Grenoble, Directeur de thèse

Monsieur, Laurent Mora
Maître de conférence à l'Université de Bordeaux, Invité
Monsieur, Olivier Adrot
Maître de conférence à l'Université de Grenoble, Invité

Table des matières

Table des figures

Notations et définitions

$X, Y, Z, \ldots$: variables aléatoires réelles

$\boldsymbol{X} = (X^1, \ldots, X^j, \ldots, X^p)$: vecteur aléatoire de dimension p et de j^{me} composante X^j

$\boldsymbol{X}^*$: transposée du vecteur $\boldsymbol{X}$

$\boldsymbol{X}^{(i)} = (X^{1,(i)}, \ldots, X^{p,(i)}), \quad i = 1, \ldots, N$: échantillon de taille N du vecteur aléatoire $\boldsymbol{X}$ de dimension p

$\boldsymbol{X}_t = (X_t^1, \ldots, X_t^p)$: processus vectoriel de dimension p

$\mathbb{X}_{t,k}$: vecteur de $(\boldsymbol{X}_{t-k}, \boldsymbol{X}_{t-k+1}, \ldots, \boldsymbol{X}_t)$

$\mathbb{X}_t$: vecteur de $(\boldsymbol{X}_0, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$

E : espérance

Var : variance

Cov : covariance

Corr : corrélation

$\mathbf{E}^X(Y)$ notée aussi $E(Y|X)$: espérance conditionnelle de Y par rapport à X

$\Gamma_{t,s} = \mathbf{Cov}(\boldsymbol{X}_t, \boldsymbol{X}_s)$ matrice de covariance

Lorsque X_t est un processus stationnaire $\Gamma_{t,s} = \mathbf{Cov}(X_t, X_s) = \gamma(t-s)$ la fonction d'auto-covariance

$\Gamma_{t,k}^X = \mathbf{E}(\mathbb{X}_{t,k}\mathbb{X}_{t,k}^*)$ et par simplification $\Gamma_{t,0}^X = \Gamma_t^X$

Pour toute variable aléatoire ϕ et tout vecteur aléatoire X_t : $\gamma_{t,k}^{X,\phi} = \mathbf{E}(\mathbb{X}_{t,k}\phi)$

F : fonction de répartition

$F_{X^j|X^1,\ldots,X^{j-1}} = F_{j|1,\ldots,j-1}$: fonction de répartition conditionnelle de X^j par rapport à $X^1, \ldots, X^{j-1}$

f : densité de probabilité

$f_{X^j|X^1,\ldots,X^{j-1}} = f_{j|1,\ldots,j-1}$: densité de probabilité conditionnelle de X^j par rapport à $X^1, \ldots, X^{j-1}$

$\overleftarrow{F}$: fonction réciproque de la fonction de répartition F

Span$\{X^i, i \in I\}$ espace linéaire engendré par $\{X^i, i \in I\}$

$S^{X^j} = \dfrac{\mathbf{Var}(\mathbf{E}(Y|X^j))}{\mathbf{Var}(Y)}$: indice de sensibilité par rapport à X^j pour le modèle $Y = \eta(\boldsymbol{X})$

$S_{t,k}^{X^j} = \dfrac{\mathbf{E}(\mathbf{E}(Y_t|\mathbb{X}_{t,k}^j)^2) - \mathbf{E}(Y_t)^2}{\mathbf{Var}(Y_t)}$: $k-$sensibilité par rapport à X^j pour le modèle $Y_t = \eta(\mathbb{X}_t)$

VARMA : Vectorial Autoregressive moving average

VAR : Vectorial Autoregressive

VARX : Vectorial Autoregressive with eXogeneous variable

VARCS : Vectorial Autoregressive Cyclo-Stationary

Remerciements

Remerciements à :

Benoit Delinchant et Olivier Adrot pour m'avoir proposé ce sujet et pour avoir encadré cette thèse. Leur disponibilité et leur patience a permis à cette thèse d'aboutir. Même si nous étions parfois en désaccord, les discussions que nous avons eues ont toujours été très enrichissantes. Merci également pour m'avoir accordé leur confiance et surtout pour avoir corrigé ce manuscrit à plusieurs reprises.

Bertrand Iooss et Stéphane Clénet pour avoir accepté d'être les rapporteurs de cette thèse. Je les remercie pour le temps passé à la relecture de ce manuscrit et pour leurs commentaires avisés. Leurs remarques et suggestions de corrections m'ont été utiles pour la rédaction finale de cette thèse.

Clémentine Prieur pour avoir accepté de présider le jury de cette thèse.

Laurent Mora pour sa participation.

Les questions de tous les membres du jury m'ont permis d'entrevoir les possibilités de futures explorations.

Je remercie également Fabrice Gamboa pour son aide et son soutien.

J'ai beaucoup apprécié que mon directeur de thèse et Edith Clavel m'aient donné l'opportunité d'encadrer des étudiants en DUT.

J'ai eu la chance d'effectuer ma thèse au G2ELab dont je remercie tous les membres pour leur gentillesse. Pendant ces 4 années de thèse j'ai eu l'occasion de rencontrer des gens formidables dont mes deux amis Gatien Kwimang et Kaustav Basu. Sans eux cette thèse aurait été sans doute encore plus dure à mener à bien.

Je tiens à remercier ma famille et particulièrement ma mère pour son soutien et ma tante pour ses relectures assidues.

Enfin, je remercie le projet ANR Fiabilité pour l'aide financière qu'ils m'ont attribuée pour ce travail.

Introduction

Aujourd'hui, la facture énergétique des bâtiments représente près de 44% de la facture globale en France, évaluée en 2012 à 69 milliards d'euros, soit une augmentation de 30,36 milliards d'euros (+11% par rapport à l'année précédente). L'efficacité énergétique des bâtiments représente donc un réel enjeu. Les bâtiments sont les principaux leviers d'action d'optimisation de l'efficacité énergétique et de réduction des émissions de CO_2 dans les villes. Selon une étude sur les bâtiments, publiée en 2012 par l'ADEME (Agence De l'Environnement et de la Maîtrise de l'Energie), le secteur du bâtiment en France est le plus gros consommateur final d'énergie. Pour comprendre comment se comporte la consommation énergétique d'un bâtiment différentes études ont été menées afin d'étudier les performances thermiques aussi bien du point de vue de la conception, de la calibration de modèle que de l'impact de changement climatique. Les performances énergétiques peuvent être optimisées pour ces différentes études en évaluant la part d'incertitude due à chacune des variables ou des paramètres qui peuvent influencer ces performances. Cette phase s'appelle l'analyse de sensibilité. C'est dans ce cadre que l'ANR "Fiabilité" a décidé d'orienter sa thématique de recherche. Elle s'est centrée sur la question fondamentale de la fiabilité des codes de simulation thermique et énergétique des bâtiments dont l'utilisation est décisive dans le processus de conception de constructions neuves ou en rénovation, dans le contexte de bâtiments à basse consommation (BBC) ou à bilan énergétique positif (BEPOS) afin d'améliorer les performances énergétiques. Lors de cette analyse on souhaite mesurer en terme d'incertitude attachée aux paramètres d'entrée l'impact sur un paramètre d'intérêt appelé sortie. On peut les classer en deux types :

- local : l'influence d'un paramètre est mesurée par la variation de la sortie autour d'une valeur nominale de ce paramètre, ce qui se traduit par la dérivée partielle de la sortie par rapport à ce paramètre.
- global : l'incertitude sur les paramètres est modélisée par une loi de probabilité a priori. Cette loi sur les entrées et les paramètres ainsi que le modèle déterminent entièrement la distribution de la sortie du modèle.

Parmi les différentes méthodes d'analyse de sensibilité, nous privilégierons la méthode globale, reposant sur le calcul des indices de sensibilité de Sobol. L'indice de Sobol d'un paramètre (ou d'un groupe de paramètres) est un indicateur statistique, d'interprétation aisée, de l'importance de ce paramètre (ou de ce groupe de paramètres) sur la variabilité d'une quantité scalaire d'intérêt, fonction de la sortie du modèle. Le calcul effectif des indices de sensibilité permet de hiérarchiser les paramètres d'entrée en fonction de leur influence sur la sortie. Un utilisateur du modèle peut alors identifier les paramètres les plus influents comme ceux sur lesquels l'incertitude doit être réduite – dans la mesure du possible – en priorité afin d'apporter une réduction de l'incertitude sur la sortie. Par ailleurs l'utilisateur a souvent un avis qualita-

tif, a priori, basé sur son intuition, sur des règles d'expertise des paramètres les plus influents d'un système physique. La confrontation de cette préconception avec les indices calculés numériquement permet de valider, ou au contraire d'invalider un modèle ou son implémentation informatique.

Les indices de Sobol peuvent se calculer de différentes façons. Dans la suite nous nous intéresserons notamment à la méthode Pick and Freeze basée sur l'échantillonnage. Celle-ci repose sur l'hypothèse fondamentale et dans la pratique le plus souvent non vérifiée d'indépendance des entrées. C'est son principal défaut. La méthode n'impose pas de forme particulière du modèle entrée-sortie. C'est un avantage important. La deuxième contrainte est qu'elle exige de pouvoir échantillonner aisément le modèle entrée-sortie. Ce n'est évidemment pas toujours le cas. Grâce au lemme de Sobol, l'indice est vu comme la covariance entre la sortie du modèle et sa réplication « Pick-Freezed ». Cette réplication est obtenue en maintenant gelée la valeur du paramètre ou de la variable d'intérêt et en échantillonnant les autres paramètres ou variables. Les répétitions de l'échantillon sont combinées pour produire par une simple loi des grands nombres un estimateur de l'indice de Sobol. Cet estimateur converge lentement et nécessite de grands échantillons.

La plupart des études en bâtiment menées dans la littérature se placent dans un cadre statique qui ne représente pas l'évolution du système. Les variables dont on souhaite étudier la sensibilité sont soit considérées à un instant donné, soit les modèles entrées-sorties ne sont pas dynamiques. Il nous est très vite apparu nécessaire de développer des méthodes qui prennent en compte à la fois la dépendance des entrées et la dimension temporelle qui elle même comporte toujours de la dépendance. Dans notre travail nous avons donc développé des méthodes pour des entrées dynamiques et dépendantes. Nous avons conservé la même définition de l'indice de Sobol mais dans une démarche différente. Nous nous sommes intéressés à ce que nous avons défini comme la mémoire utile au calcul de la sensibilité. Cette mémoire utile est liée physiquement à l'inertie thermique des différentes composantes du système. La première approche naïve consiste à calculer ce que l'on appellera la sensibilité instantanée. On regarde la sensibilité de la sortie à chaque instant par rapport à la variable d'entrée X_t^1 à l'instant t :

$$S_t^{X^1} = \frac{\mathbf{Var}\left(\mathbf{E}(Y_t|X_t^1)\right)}{\mathbf{Var}(Y_t)}$$

Remarquons que ce point de vue naïf est plus compliqué déjà qu'il peut le sembler à première vue. En effet la mémoire de la sensibilité va dépendre de la mémoire propre à l'entrée et d'une autre mémoire qui dépend de la relation entrée sortie qui n'est pas toujours instantanée, par exemple si la sortie est de type récursif.

Donc cette sensibilité instantanée n'est pas en général la mieux appropriée dans un cadre dynamique car le processus de sortie à l'instant t dépend de ses instants passés Y_{t-k} et par la même des instants passés du processus d'entrée X_{t-k}^1. Il est plus judicieux alors de calculer la sensibilité par rapport à $(X_t^1, \ldots, X_{t-k}^1)$. On définit ce que l'on appellera les $k-$sensibilités :

$$S_{t,k}^{X^1} = \frac{\mathbf{Var}\left(\mathbf{E}(Y_t|X_t^1, \ldots, X_{t-k}^1)\right)}{\mathbf{Var}(Y_t)}$$

On comprend alors que lorsque l'on souhaite calculer la sensibilité de X_t sur Y_t il faut prendre en compte tout le passé de X d'où la définition de l'indice POPSI (Projection On The Past

Sensitivity Index) :

$$S_{t,k}^{X^1} = \frac{\mathbf{Var}\left(\mathbf{E}(Y_t|X_t^1,\ldots,X_0^1)\right)}{\mathbf{Var}(Y_t)}$$

Afin de s'affranchir de l'hypothèse d'indépendance de la méthode Pick and Freeze nous présentons une méthode dans deux cas. Dans un premier temps nous proposons d'étudier le cas Gaussien, le plus utilisé dans les problèmes physiques du type de ceux du bâtiment. Nous proposons de réécrire le modèle $Y_t = \eta(\boldsymbol{X}_t, \boldsymbol{Z}_t)$ sous une nouvelle forme $Y_t = g(\boldsymbol{X}_t, \boldsymbol{W}_t)$ avec $\boldsymbol{W}_t$ un processus indépendant de $\boldsymbol{X}_t$. A partir de la structure temporelle des processus Gaussien les variables en entrée $(\boldsymbol{X}_t, \boldsymbol{Z}_t)$ peuvent se réécrire $(\boldsymbol{X}_t, \tilde{\boldsymbol{X}}_t + \boldsymbol{W}_t)$. Un algorithme est proposé dans la partie 3 pour calculer le processus $\boldsymbol{W}_t$.

Dans le cas non Gaussien nous partons d'un résultat concernant les modèles statiques. L'idée est de construire une transformation permettant de passer de variables $(X^1,\ldots,X^p)$ de loi quelconque à $(U^1,\ldots,U^p)$, p variables indépendantes de loi uniforme. On montrera comment utiliser cette transformation pour définir un nouveau modèle $Y = \tilde{\eta}(U)$. Cette méthode très puissante car très générale peut devenir lourde dès que la dimension temporelle (la mémoire "utile") est grande.
Qualitativement, les modèles d'entrée peuvent ne pas être gaussiens. Par exemple les lois marginales peuvent être bi-modales, fortement asymétriques, être à support borné ou au contraire avec des queues de probabilité très lourdes. Il est malheureusement très difficile d'obtenir des modélisations temporelles simples avec des lois marginales et des covariances fixées par avance. En dimension $p > 2$, la recherche de solutions sous contraintes de type auto-régressif est un problème qui n'admet pas de solution en général mais qui admet très probablement de manière très générale des solutions approchées de bonne qualité. Nous avons repris l'idée des distributions de Johnson (qui peuvent être bi-modales, à support borné, ...) et montré que l'on peut étendre la méthode Pick and Freeze à ces processus, ce qui ouvre un chemin prometteur à des calculs de sensibilité pour des systèmes dynamiques non gaussiens.

Toutes ces méthodes de calcul de sensibilité ont été appliquées à un bâtiment test dans la partie 3. Cette partie propose des méthodes pouvant s'appliquer à d'autres cas. Le manque de données nous a amenés à considérer les données par mois. Nous avons considéré un mois été et un mois hiver pour construire les modèles entrées-sortie. Différents modèles pour les entrées sont proposés à qui sont associés des simulateurs dans le but d'appliquer la méthode Pick and Freeze. Les différentes sensibilités ont été calculées sur ces modèles. La faiblesse des données et la complexité de celles-ci (plusieurs saisonnalités, changement de régime de la variable de chauffage...) influencent la qualité des modèles et par la même, les résultats sur la sensibilité.

Le plan de notre travail est le suivant :

- La première partie présente les outils probabilistes et statistiques adaptés à un cadre d'entrées statiques. Après un rappel sur la construction des méta-modèles entrée-sortie, nous présenterons la façon de propager des incertitudes et de les quantifier dans le cadre d'entrées indépendantes ; nous présenterons également la méthode que nous avons développée pour des entrées dépendantes.
- La deuxième partie expose les outils utilisés dans un cadre dynamique. Nous présenterons des méta-modèles utilisés dans ce cadre et développerons une méthode adaptée au calcul d'indices de sensibilité pour des entrées dépendantes et dynamiques.

— La dernière partie propose tout d'abord une présentation de la problématique de la sensibilité dans les échanges énergétiques dans un bâtiment, puis une application à un bâtiment test des méthodes développées précédemment, en particulier celle des indices de sensibilité dynamiques.

Première partie

Outils probabilistes et statistiques

Chapitre 1

Méta-modèles et sensibilité

1.1 Méta-modèles statistiques

1.1.1 Méta-modèles

Les modèles mathématiques sont usuellement la traduction en langage mathématique de modèles physiques. Ceux-ci sont a priori la traduction d'une réalité physique. Il en est ainsi de la modélisation des températures à l'intérieur d'un bâtiment soumis à la contrainte intérieure d'un système de chauffage régulé et à des contraintes extérieures (comme la température de l'air entourant le bâtiment et la quantité de chaleur apportée par les visiteurs), situation que nous allons rencontrer dans la deuxième partie de cette thèse. Ces modèles peuvent être utilisés à des fins de prévisions ou de gestion par exemple. Les observations réelles faites sur le bâtiment et les autres variables concernées peuvent être en nombre très limité, en tout cas insuffisant pour analyser la situation. Ceci devient une évidence s'il s'agit de faire le modèle d'un bâtiment dont on projette la construction. On commence plutôt par des modélisations numériques permettant de résoudre les lois physiques. Ces modèles sont parfois lourds ce qui nous conduit à travailler avec des modèles physiques plus légers (par exemple des circuits électriques équivalents). Il est en général très difficile de traduire certaines formes de non linéarité ou bien des phases transitoires entre régimes permanents.

Parmi toutes les entrées et les paramètres qui interviennent dans le modèle que l'on construit, certains peuvent être assez mal connus (beaucoup d'incertitudes sur leur valeur), mais n'influencer que très peu la sortie du modèle (peu de variations). Au contraire, d'autres peuvent être connus avec peu d'incertitudes, mais avoir un énorme impact sur une sortie. Identifier les entrées et les paramètres influents est par conséquent primordial pour mieux cerner le comportement du modèle ou détecter des anomalies : c'est l'analyse de sensibilité. Lorsque l'on s'intéresse aux incertitudes autour des valeurs nominales recherchées, plus précisément au transfert d'incertitudes entre des variables d'entrée relativement faciles d'accès et une variable de sortie d'accès plus difficile, le modèle physique peut être d'utilisation trop compliquée, ou prenant trop de temps de calcul. Dans ce cas on peut recourir à des modèles mathématiques plus faciles à étudier et permettant aussi de guider le plan de nouvelles expériences. Ces modèles sont appelés méta-modèles, le terme méta signifiant ici que l'on a conscience d'un certain

décalage avec la réalité physique. S'agissant d'incertitude, ces modèles seront quasi systématiquement des modèles stochastiques. Nous verrons que souvent par le passé le travail des ingénieurs se fondait sur des notions en apparence déterministes, c'est le cas de différentes définitions anciennes de la sensibilité mais très rapidement on a abouti à replacer ces notions dans un cadre aléatoire.
La nécessité de disposer d'un ensemble assez important de modèles pour pouvoir les adapter à la fois aux données mais aussi à des contraintes diverses fait que le modèle stochastique va dépendre d'un certain nombre de paramètres à estimer, non connus a priori avant toute expérience. Une collection de tels modèles stochastiques est appelée modèle statistique et formellement représenté par une famille de probabilités P_θ définies sur le même espace de probabilité, le paramètre $\theta \in \Theta$, θ pouvant être une partie d'un espace euclidien ou d'un espace fonctionnel de dimension infinie.

Un méta modèle est donc comme tout modèle une certaine approximation de la réalité physique. La qualité de cette approximation dépend évidemment de façon décisive de la qualité, de la précision et du nombre, des données. Elle dépend aussi du choix du modèle choisi suivant des critères mathématiques. La construction d'un modèle statistique doit répondre aux premiers impératifs suivants, étroitement reliés :

- Ne pas comporter trop de paramètres à estimer à partir des données (principe de parcimonie)
- Réaliser un bon compromis biais/variance ou en termes plus généraux un bon compromis ajustement/ prédiction. Un modèle surajusté (trop "voisin" des données) fera de mauvaises prédictions en sous estimant la variabilité intrinsèque des données.

En pratique ces principes se traduisent par des critères mathématiques concernant la sélection de modèles à l'intérieur d'une classe donnée a priori. L'ensemble de ces critères et de pratiques statistiques sera résumé dans un langage peu formalisé au paragraphe suivant. En ce qui concerne cette thèse, l'ensemble de ces considérations qui sont au coeur de la statistique moderne sera en particulier appliqué à des classes de modèles particuliers de séries chronologiques.

1.1.2 Construction de méta-modèles statistiques entrée-sortie : cas statique

Nous allons nous intéresser d'abord au cas le plus simple et le plus développé en ce qui concerne la construction de modèles sur lesquels vont reposer les études de sensibilité qui sont l'objet de ce travail. On suppose disposer d'un jeu de données dites d'observation représentant un modèle entrée-sortie. Les données sont donc un ensemble de couples $(\boldsymbol{X}^{(i)}, Y^{(i)}), i = 1, \ldots, N$. On ne cherche pas dans ce paragraphe à se servir de ces données pour en fabriquer d'autres ou pour construire un plan d'expérience efficace. On cherche à construire un méta-modèle qui soit cohérent avec la théorie statistique quitte à le simplifier éventuellement dans une deuxième étape.

Définition 1. *Un méta-modèle statistique entrée-sortie sera donné par une relation :*

$$Y = \eta(\boldsymbol{X}, \theta) \ \ , \theta \in \Theta$$

où $\boldsymbol{X}$ est un vecteur aléatoire de dimension finie, $\boldsymbol{X} \in \mathbb{R}^p$, et θ un paramètre, $\theta \in \Theta$, Θ de dimension finie ou infinie, Y est une variable aléatoire réelle donc : $\eta : \mathbb{R}^p \times \Theta \to \mathbb{R}$

Phase 1 : construction

La phase 1 est exploratoire pour trouver des relations possibles entre Y et $\boldsymbol{X}$. Par exemple si $\boldsymbol{X} = (X^1, \ldots, X^p)$, on peut étudier a priori un modèle additif non paramétrique d'ordre 1 puis un modèle d'ordre 2 soit :

$$Y = \sum_{j=1}^{p} h_j(X^j)$$

$$Y = \sum_{j=1}^{p} h_j(X^j) + \sum_{1 \leq i \neq j \leq p}^{p} h_{i,j}(X^i, X^j)$$

la forme des fonctions $h_j, h_{i,j}$ (éventuellement soumises à contraintes) donnant des indications sur η.

Phase 2 estimation des paramètre du modèle (identification paramétrique)

On choisit une famille $\eta(\boldsymbol{X}, \theta)$ avec $\theta \in \Theta$. La dimension de Θ (souvent mesurée en termes d'entropie métrique) mesure la complexité du modèle. Le choix de θ (et de η) va toujours répondre aux impératifs exposés en 1. Si $dim(\theta)$ est trop importante, le modèle est sur-paramétré, il est choisi "trop près" des données : il est sur-ajusté. Le modèle sera un bon descripteur des observations mais un mauvais prédicteur : si on l'utilise sur un autre échantillon $(\boldsymbol{X}'^{(i)}, Y'^{(i)})_{i=1\ldots N}$, distinct de l'échantillon d'observation de départ, la relation $Y''^{(i)} = H(\boldsymbol{X}'^{(i)}, \theta)$ définissant la prédiction de $Y'^{(i)}$ on aura $\|Y''^{(i)} - Y'^{(i)}\|$ trop importante. Le but premier des méthodes statistiques est de trouver un compromis entre une bonne description des données de base et une bonne prédiction des données à venir. Pour choisir θ, il faut un critère souvent appelé critère de risque ou contraste notée R, qui est en fait une pseudo-distance sur Θ. Un bon estimateur $\hat{\theta}$ de la "vraie" valeur θ_0 de θ (valeur optimale inconnue) est une des valeurs telles que :

$$\hat{\theta} = \underset{\theta \in \Theta}{argmin} \; R(\theta_0, \theta(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}))$$

où $\hat{\theta}(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}) : \mathbb{R}^{Np} \to \Theta$.

Comme θ_0 est inconnu, cette définition n'est pas utilisable directement. On approche le contraste par une suite stochastique qui elle ne dépend pas de θ_0. Les 2 exemples les plus classiques sont les moindres carrés (MC) et la vraisemblance (ou son logarithme). Pour les moindres carrés, le risque vaut :

$$\|\theta_0 - \hat{\theta}\|^2$$

si $\Theta \subset \mathbb{R}^p$ et il est approché par :

$$\frac{1}{N} \sum_{i=1}^{N} |Y^{(i)} - \theta \boldsymbol{X}^{(i)}|^2$$

dans le cas d'un modèle du type : $Y^{(i)} = \theta \boldsymbol{X}^{(i)} + \varepsilon^{(i)}$ où $\varepsilon^{(i)}$ est un bruit blanc, $\mathbf{E}(\varepsilon^{(i)}) = 0$, $\mathbf{E}((\varepsilon^{(i)})^2) = \sigma^2 \geq 0$.
Dans le cas de la vraisemblance, le risque est la fonction d'information de Kullback ([26] pour ce type de modèle) approchée par le logarithme de la vraisemblance. Mais cette méthode de choix ne permet pas en général de trouver le compromis souhaitable, elle amène automatiquement à un sur-paramétrage. On utilise alors des procédures de sélection de modèle (ou de sélection de l'ordre d'un modèle). On pose :

$$\Theta = \sqcup_{k=1}^{K} \Theta^k \text{ avec } \Theta^k \subset \Theta^{k+1} \text{ et } K \leq \infty$$

Si $K < \infty$, on parlera de sélection de l'ordre. Si $K = \infty$, on parlera de sélection tout court. La suite Θ^k définit donc des modèles de complexité croissante.
On garde les mêmes concepts et notations. On se donne une fonction de risque $R(\theta_0, \theta)$ approchée par une suite de fonctions $L_N(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}, \theta)$ (comme la somme des carrés ou l'opposé du logarithme de la vraisemblance divisé par N). Maximiser L_N en θ amènerait à prendre θ dans le plus grand des espaces Θ^k. On introduit donc une pénalisation $pen(N, k)$ qui est une fonction croissante de k (et le plus souvent décroissante de N) et on va estimer θ par $\hat{\theta}_N$.

$$\begin{aligned}\hat{\theta}_N &= \underset{k}{argmin} \ \underset{\theta \in \Theta^k}{min} L_N(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}, \theta) + pen(N, k) \\ &= \underset{k}{argmax} \ \underset{\theta \in \Theta^k}{max}(-L_N(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}, \theta) - pen(N, k))\end{aligned}$$

Cette formulation est celle utilisée pour la vraisemblance.
Des pénalisations classiques (Akaike [55], Schwartz [5], cas bayésien, ...) sont du type $\phi(N, k)$ pour les problèmes du choix de l'ordre du modèle ($K < \infty$). Nous ne détaillerons pas ici un formalisme général pour la sélection de modèles pour $K = \infty$. Nous y reviendrons plus loin sur des exemples concernant la sensibilité.

Il existe une autre lecture, très classique, du compromis "ajustement/prédiction". C'est le compromis "biais/variance". Si Θ est une autre partie d'un espace normé, de dimension quelconque et si θ_0 est la "vraie" valeur du paramètre, lorsque l'on choisit $\hat{\theta}$ pour estimer θ on a alors un biais :

$$b(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta_0)$$

et une variance d'erreur :

$$\mathbf{E}(\|\hat{\theta} - \theta_0\|^2) = \|b(\hat{\theta})\|^2 + \mathbf{Var}(\hat{\theta})$$

de sorte que :

$$\mathbf{Var}(\hat{\theta}) = \mathbf{E}(\|\hat{\theta} - \mathbf{E}(\hat{\theta})\|^2)$$

Un biais trop fort signifie un mauvais ajustement. k est trop faible. A contrario un biais trop faible équivaut à un sur-ajustement et donc à une mauvaise performance de $\hat{\theta}$ sur un autre échantillon que celui qui a servi à l'estimation.

Phase 3 : Validation

Il s'agit de s'assurer autant que possible que le modèle est de bonne qualité. Il y a deux situations :

— Les données sont assez nombreuses pour ne pas avoir à les utiliser toutes pour la partie apprentissage, c'est-à-dire ici l'estimation du modèle. On divise les données en 2 paquets. Un premier est utilisé pour ce qui vient d'être dit de l'estimation (phase 2). Le second est utilisé pour mesurer la qualité grâce à un critère fixé a priori de validation, comme par exemple l'erreur globale de prévision. Sur le nouvel échantillon $(\boldsymbol{X}'^{(i)}, Y'^{(i)})_{i=1,\ldots,M}$ on va calculer :

$$\sum_{i=1}^{M} |Y'^{(i)} - \eta(\hat{\theta}(\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(N)}))|^2$$

Si cette erreur est trop forte, le modèle ne sera pas validé.

— Les données sont rares ou peu nombreuses. On doit les utiliser toutes pour la partie estimation. On procède alors par des méthodes devenues standards qui sont la validation croisée et le jackknife qui en est un cas particulier. On se fixe un nombre petit, m, entre 1 (jackknife) et 5. On considère tous les échantillons de dimension $N - m$ pour lesquels on a ôté m variables. Pour chacun de ces échantillons, noté par l'indice j, on calcule $\hat{\theta}^j$. Puis on calcule :

$$CV(m) = \sum_{j} \sum_{\substack{i=1 \\ i \neq i_1,\ldots,i_m}}^{N} \|\eta(\hat{\theta}, \boldsymbol{X}^{(i)}) - \eta(\hat{\theta}^j, \boldsymbol{X}^{(i)})\|^2$$

où $i_1, \ldots, i_m$ sont les indices ôtés correspondant à j. Ce critère a des propriétés mathématiques intéressantes pour de très larges classes de modèles. Sur certains modèles, en particulier linéaires, on verra sur des exemples, l'utilisation du bootstrap ([36]) à des fins de validation.

Chapitre 2

Propagation des incertitudes et sensibilité

2.1 Propagation de l'incertitude

Les modèles physiques étudiés ici ont une expression mathématique du type $y = \eta(\boldsymbol{x})$ où $\boldsymbol{x} = (x^1, \ldots, x^p)$ sera considéré dans cette partie comme un vecteur aléatoire à p composantes réelles. Au voisinage d'un point de fonctionnement $\boldsymbol{x}^0$ la variable y va se trouver, si le système a une certaine stabilité, dans le voisinage de y^0. La problématique est la suivante : on connaît $\boldsymbol{x} = (x^1, \ldots, x^p)$ avec une certaine incertitude, différente suivant les x^i , $i = 1, \ldots, p$. Comment cette incertitude va-t-elle se propager sur y ?

Ainsi formulé le problème a déjà une formulation probabiliste puisque toute traduction des incertitudes se fera en terme de probabilités. Néanmoins, pour des raisons historiques, l'introduction de l'aléatoire dans le domaine de l'ingénierie s'est faite assez récemment et donc des formulations en apparence déterministes demeurent.
Nous allons les rappeler en premier en ce qui concerne la sensibilité, notion qui vise à mesurer l'importance de la construction de l'incertitude sur chacune des variables x^i dans l'incertitude observée ou attendue sur l'influence de y.

Remarque 1. *Les termes de variable et de paramètre sont souvent utilisés de façon peu différenciée dans la littérature concernant la sensibilité. Le modèle $y = \eta(\boldsymbol{x})$ dépend de fait de paramètres θ dont la valeur est incertaine, ce qui sera le cas le plus souvent pour des méta-modèles de type statistique. Cette incertitude sur les paramètres sera transmise à y. Dans ce cas les paramètres seront traités éventuellement en ce qui concerne la partie sensibilité comme des variables.*

Dans ce qui suit, on s'intéresse à la sortie d'un modèle $y = \eta(x^1, \cdots, x^p)$ où les paramètres d'entrées $\boldsymbol{x} = (x^1, \cdots, x^p)$ sont déterministes ou aléatoires. Pour mesurer l'importance d'une variable d'entrée sur la sortie y, il faut disposer d'un indicateur d'importance qui peut être qualitatif (par ex. le plan de Morris [87]) ou quantitatif (par ex. l'indice de Sobol [109]).

2.2 Indices de sensibilité de criblage

La première méthode appelée screening fournit une information qualitative à faible coût de calcul sur l'importance d'un facteur. Cette méthode est typiquement utilisée comme première étape d'une analyse de sensibilité lorsque le nombre de facteurs est grand, et que l'on veut exhiber rapidement un groupe de facteurs influents. Elle propose une information globale mais dépend de l'échantillonnage et ne permet pas de quantifier les interactions. L'approche la plus utilisée et que nous allons exposer est celle de Morris : OAT (One-At-a-Time).

Soit $\boldsymbol{x}^{(1)}$ un jeu de valeurs des entrées choisi en respectant leurs distributions marginales (notion évidemment probabiliste). On choisit un second jeu de valeurs $\boldsymbol{x}^{(2)}$ de la même manière et on détermine le vecteur pas d'incrément $\boldsymbol{\Delta} = \boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)} = \{\Delta_1, \ldots, \Delta_p\}$. Puis, après avoir évalué $\eta(\boldsymbol{x}^{(1)})$, on évalue la fonction en modifiant une valeur à la fois. Ainsi, à la i-ième itération, on évaluera η pour le jeu de valeurs suivant : $\boldsymbol{x}^{(i)} = \{x^{1,(1)}, \ldots, x^{i,(1)} + \Delta_i, x^{i+1,(1)}, \ldots, x^{p,(1)}\}$. On définit alors l'effet élémentaire du i-ième facteur en un point $\boldsymbol{x}$ par :

$$d_i^{(1)}(\boldsymbol{x}) = \frac{\eta(\boldsymbol{x}^{(i)}) - \eta(\boldsymbol{x}^{(1)})}{\Delta_i}. \tag{I.1}$$

On détermine ainsi p effets élémentaires valables au voisinage d'un point $\mathbf{x}^{(1)}$. Cela donne alors une information locale. En répétant le calcul N fois en choisissant de nouveaux points dans l'espace des paramètres tout en veillant à bien couvrir l'espace et à respecter les distributions marginales on obtient $N \times p$ effets élémentaires fournissant une information *globale*. L'analyse des moyennes μ_i et écart-types σ_i des effets élémentaires renseigne sur l'importance de chaque facteur sur la sortie étudiée. Ces indicateurs valent :

$$\begin{aligned} m_i &= \frac{1}{N}\sum_{j=1}^{N} |d_i^{(j)}| \\ \sigma_i^2 &= \frac{1}{N-1}\sum_{j=1}^{N} (d_i^{(j)} - m_i)^2 \end{aligned} \tag{I.2}$$

L'interprétation classique est à manier avec précaution et est du type :

	σ_i faible	σ_i élevé
m_i faible	négligeable	non linéarité et/ou interactions
m_i élevé	influente	non linéarité et/ou interactions

Si la sortie est scalaire, pour faciliter l'analyse, on peut tracer un graphique dans le plan de Morris représentant les points de coordonnées (m_i, σ_i). Dans le cas contraire, la distance suivante peut être employée pour analyser l'influence d'un paramètre :

$$\delta_i = \sqrt{m_i^2 + \sigma_i^2}. \tag{I.3}$$

On peut remarquer que cette méthode est d'abord une approximation en chaque point d'échantillonnage du gradient de la sortie $y = \eta(\boldsymbol{x})$ en fonction des entrées $x^i, i = 1, \ldots, p$.

$$S(\boldsymbol{x}) = \nabla\eta(\boldsymbol{x}) = \left(\frac{\partial\eta(\boldsymbol{x})}{\partial x^1}, \ldots, \frac{\partial\eta(\boldsymbol{x})}{\partial x^p}\right) \tag{I.4}$$

Le gradient offre une sensibilité locale car il dépend du point de calcul $\boldsymbol{x}$.

Il existe plusieurs méthodes pour obtenir un gradient. Lorsque l'on possède une expression de η composée d'expressions mathématiques simples, utilisant des fonctions classiques dont les dérivées sont bien connues, il est facile d'obtenir le gradient en appliquant des théorèmes de composition de dérivées. Pour des formulations plus complexes, il est avisé d'utiliser des propriétés mathématiques plus sophistiquées (utilisation de méthodes adjointes [34]). Nous ne disposons cependant pas toujours de propriétés mathématiques permettant d'obtenir directement les dérivées. Dans ce cas, il est possible de formuler des modèles sous la forme de codes informatiques en langage C ou Java par exemple. En utilisant la dérivation de code, nous pouvons obtenir, dans de nombreux cas, le jacobien du modèle. La Dérivation Automatique (DA) exploite le même processus que la compilation de code de programmation. Les dérivées de chaque opération élémentaire sont combinées selon la règle de la chaîne du calcul différentiel afin d'obtenir la dérivée d'une instruction plus complexe. Cette technique fournit un moyen efficace pour calculer des gradients. Des solutions adaptées aux modélisateurs existent à partir de différents langages tels que C [39] ou JAVA [93]. Dans certains cas, la non dérivabilité de fonctions en certains points peut perturber l'algorithme d'optimisation. Il s'agit parfois de non dérivabilités introduites dans la modélisation et ne présentant pas un caractère physique comme dans le cas de la dérivation d'un algorithme de recherche de minimum d'une fonction qui peut " sauter " d'un minimum local à un autre selon l'évolution de la fonction.

Sobol et al. [110] et Kucherenko et al. [75] proposent de prendre la moyenne de l'indice local. On considère la fonction différentiable η et $\boldsymbol{x} = (x^1, \ldots, x^p)$ le vecteur d'entrée défini sur l'hypercube unité ; c'est-à-dire $0 \leq x^i \leq 1$:

$$\mathbf{M}_i = \int \frac{\partial\eta(\boldsymbol{x})}{\partial x^i}\mathrm{d}\boldsymbol{x} \tag{I.5}$$

De la même manière on considère la variance :

$$\mathbf{\Sigma}_i^2 = \int (\frac{\partial\eta(\boldsymbol{x})}{\partial x^i} - \mathbf{M}_i)^2\mathrm{d}\boldsymbol{x} \tag{I.6}$$

En remarquant que :

$$\mathbf{\Sigma}_i^2 = \int \left(\frac{\partial\eta(\boldsymbol{x})}{\partial x^i}\right)^2 \mathrm{d}x - \mathbf{M}_i^2$$

On définit G_i un indice global de sensibilité :

$$\mathbf{G}_i = \int \left(\frac{\partial\eta(\boldsymbol{x})}{\partial x^i}\right)^2 \mathrm{d}\boldsymbol{x} = \mathbf{\Sigma}_i^2 + \mathbf{M}_i^2 \tag{I.7}$$

Ces trois mesures $(\mathbf{G}_i, \mathbf{M}_i, \mathbf{\Sigma}_i)$ définissent ce que l'on appelle l'indice DGSM (Derivative Global Sensitivity Measures).

Le choix du plan d'expérience est dans tous ces cas essentiel.
Un estimateur classique est :

$$\hat{\mathbf{G}}_i = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\partial \eta(\boldsymbol{x}^{(k)})}{\partial x^i}\right)^2 \tag{I.8}$$

où $x^{(k)}$ pour $1 \leq k \leq N$ est un point de l'espace des paramètres.

Le calcul de cet indice fait appel à des méthodes de calcul d'estimateur de l'espérance (I.5) et de la variance (I.6). En optimisant le choix du plan d'expérience on peut diminuer considérablement le temps de calcul de ces indices, par exemple en utilisant une méthode de Quasi Monte Carlo (QMC).

2.3 Indices de sensibilité probabilistes

2.3.1 Décomposition de la variance sur des sous espaces orthogonaux

Ces indices sont fondés sur l'interprétation de l'espérance conditionnelle en terme de prévision.

Considérons des variables aléatoires : $X^1, \ldots, X^p$ centrées de distribution μ connue.

Considérons l'espace de Hilbert $L^2(\mu)$ engendré par toutes les fonctions de ces variables du type $\phi(X^1, \ldots, X^p)$ de carré $\mu-$intégrable.

$L^2(\mu)$ est muni du produit scalaire de la covariance :

$$< \phi, \psi >= \mathbf{Cov}(\phi(X^1, \ldots, X^p), \psi(X^1, \ldots, X^p))$$

$H = L^2(\mu)$ est un espace de dimension infinie, en identifiant toutes les fonctions qui diffèrent par une constante.

Considérons le modèle : $Y = \eta(X^1, \ldots, X^p)$, $Y \in H$.
La meilleure prévision de Y par une fonction de la seule variable X^i est l'espérance conditionnelle : $\mathbf{E}(Y|X^i)$.

Soit l'espace H^i engendré par une seule variable X^i. H^i est donc le sous espace de H des fonctions du type $\phi(X^i)$. La projection de Y sur H^i, notée $\Pi_{X^i}(Y)$, est ici $\mathbf{E}(Y|X^i)$. C'est la fonction dépendant uniquement de X^i qui approche le mieux la variable Y. L'erreur de prévision est caractérisée par l'espérance de son carré :

$$\mathbf{E}\left(\left(Y - \mathbf{E}(Y|X^i)\right)^2\right)$$

Les règles simples de calcul sur les espérances conditionnelles (par exemple [26]) donnent :

$$\mathbf{E}\left(\left(Y - \mathbf{E}(Y)|X^i)\right)\right)^2 = \mathbf{Var}(Y) - \mathbf{Var}\left(\mathbf{E}(Y|X^i)\right)$$

et donc plus $\mathbf{Var}\left(\mathbf{E}(Y|X^i)\right)$ est grande, meilleure est la prédiction.

Par conséquent, $\mathbf{Var}(\mathbf{E}(Y|X^i))$ serait la variance de la sortie Y si celle-ci était fonction uniquement de X^i. Plus cette quantité est proche de $\mathbf{Var}(Y)$, plus le facteur X^i explique la variance de Y. Autrement dit, dans un tel cas le facteur X^i est influent.

On définit alors les indices de sensibilité.
L'indice de Sobol du premier ordre est défini par :

$$S^{X_i} = \frac{\mathbf{Var}(\mathbf{E}(Y|X^i))}{\mathbf{Var}(Y)} \tag{I.9}$$

On peut faire le même raisonnement sur les espaces engendrés par plusieurs variables. La sortie est influencée à travers l'interaction des variables qui engendrent ces espaces. Pour ne garder l'effet que de l'interaction, on lui soustrait les effets d'ordres inférieurs. Par exemple l'indice d'ordre 2 est défini pour $i \neq j$ par :

$$S^{X_i,X_j} = \frac{\mathbf{Var}(\mathbf{E}(Y|X^i,X^j)) - \mathbf{Var}(\mathbf{E}(Y|X^i)) - \mathbf{Var}(\mathbf{E}(Y|X^j))}{\mathbf{Var}(Y)} \tag{I.10}$$

2.3.2 Indices de Sobol pour des entrées indépendantes et décomposition de Hoeffding

L'espace de Hilbert H est formé de variables d'espérance nulle dites centrées. Si Y n'est pas elle même centrée, sa projection sur l'espace des constantes est $\mathbf{E}(Y)$. Dans ce qui suit la variance étant invariante par centrage on pourra considérer Y comme centrée sans ne rien changer aux indices de sensibilité. L'indépendance implique alors H^i orthogonal à H^j pour $i \neq j$. En effet :

$$\mathbf{E}\left(g(X^i)h(X^j)\right) = \mathbf{E}\left(g(X^i)\right)\mathbf{E}\left(h(X^j)\right) = 0$$

Soit H^{ij} l'espace des fonctions du type : $g(X^i, X^j)$.
Si l'on pose :

$$H^{i,j} = H^i \oplus H^j \oplus K^{i,j}$$

l'espace $K^{i,j}$ sera interprété comme l'espace des interactions en X^i et X^j. Ce terme provient du modèle linéaire en statistique où la notion d'interaction est classique [3].
On a donc :

$$K^{i,j} = H^{i,j} \ominus H^i \ominus H^j$$

Ceci amène à la décomposition de la variance du type ANOVA et à la décomposition de Hoeffding. La décomposition ANOVA (acronyme pour ANalysis Of VAriance) telle que définie par [108] peut s'énoncer comme suit. Cette décomposition Eq. (I.13) permet le calcul par projection d'espérance conditionnelle du type $\mathbf{E}(Y|X^{i_1},\ldots,X^{i_p})$. Cela n'est vrai que si les variables d'entrées sont indépendantes [89].

Proposition 1. *Décomposition de Hoeffding pour des variables indépendantes et de loi uniforme.*
Soit $\eta : [0,1]^p \to \mathbb{R}$, $\int_{[0,1]^p} \eta^2(\boldsymbol{x})\mathrm{d}x < \infty$, *alors* η *admet une unique décomposition de la forme :*

$$\eta(\boldsymbol{x}) = \eta_0 + \sum_{i=1}^{p} \eta_i(X^i) + \sum_{1\leq i<j\leq p}^{k} \eta_{ij}(X^i,X^j) + \cdots + \eta_{1\ldots p}(X^1,\ldots,X^p) \tag{I.11}$$

sous les contraintes :

— $\eta_0 = \int_{\Omega^p} \eta(\boldsymbol{x})\mathrm{d}x$ *est une constante*

— $\int_0^1 \eta_{i_1\dots i_s}(x^{i_1},\dots,x^{i_s})\mathrm{d}x^{i_t} = 0 \quad si\ 1 \le t \le s$

Toutes les fonctions intervenant dans la décomposition sont orthogonales :

$$\int_{\Omega^p} \eta_{i_1\dots i_s}\eta_{j_1\dots j_l}\mathrm{d}x = 0 \quad pour \quad (i_1,\dots,i_s) \neq (j_1,\dots,j_l) \tag{I.12}$$

Donc si : $H = \eta(X^1,\dots,X^p)$

$$\eta(X^1,\dots,X^p) = \sum_{J\in P} \eta_J(X^{(J)})$$

où P est l'ensemble des parties de $\{1,\dots,p\}$ et $X^{(J)} = \{X^i, i \in J\}$

En élevant au carré et en prenant l'espérance de la décomposition, on obtient :

$$D = \sum_{i=1}^{p} D_i + \sum_{1\le i<j\le d} D_{ij} + \dots + D_{1\dots d}. \tag{I.13}$$

D'où, la définition des indices de sensibilité de Sobol basés sur la variance :

$$S^{X_{i_1}\dots X_{i_s}} = \frac{D_{i_1\dots i_s}}{D}$$

qui mesurent la part de la variance de Y due aux interactions d'ordre s des variables $\{X^{i_1},\dots,X^{i_s}\} = X^{(J)}$.

Cette décomposition conduit à une propriété importante : la somme des indices de tous les ordres est égale à 1 :

$$\sum_{i=1}^{p} S^{X^i} + \sum_{1\le i<j\le d} S^{X^iX^j} + \dots + S^{X^1\dots X^d} = 1 \tag{I.14}$$

On peut ainsi définir $2^p - 1$ indices de sensibilité.

Définition 2. *En pratique, uniquement quelques indices de sensibilité sont recherchés (pour leur utilité) :*

— *Indice de Sobol du premier ordre défini par :*

$$S^{X_i} = \frac{\mathbf{Var}(\mathbf{E}(Y|X^i))}{\mathbf{Var}(Y)} = \frac{D_i}{D} \tag{I.15}$$

— *Indice de Sobol du second ordre pour $i \neq j$:*

$$S^{X^i,X^j} = \frac{\mathbf{Var}(\mathbf{E}(Y|X^i,X^j)) - \mathbf{Var}(\mathbf{E}(Y|X^i)) - \mathbf{Var}(\mathbf{E}(Y|X^j))}{\mathbf{Var}(Y)} = \frac{D_{ij}}{D} \tag{I.16}$$

— *Indice de Sobol total d'ordre i :*

$$S_{tot}^{X^i} = S^{X^i} + \sum_{\substack{j=1 \\ j\neq i}}^{d} S^{X^i,X^j} + \cdots + S^{X^1,\ldots,X^i,\ldots,X^d} = \frac{\sum_{i\in\{i_1,\cdots,i_s\}} D_{i_1\cdots i_s}}{D} \tag{I.17}$$

Remarque 2. *Lien entre le cas déterministe et le cas probabiliste :*
Pour $\frac{\partial\eta(\boldsymbol{x})}{\partial X^i} \in L^2$ *on a :*

$$S_{tot}^{X^i} \leq \frac{G_i}{\pi^2 D} \tag{I.18}$$

où $D = \mathbf{Var}(\eta(\boldsymbol{X}))$ *et* G_i *l'indice DGSM ([76])*

Cas de variables indépendantes et de loi quelconque

Si X^i a pour fonction de répartition F_i, on pose $U^i = F_i(X^i)$. U_i est alors uniforme.
$X^i = \overleftarrow{F}_i(U^i)$. On pose $\tilde{\eta}(U^1,\ldots,U^p) = \eta(\overleftarrow{F}_1(U^1),\ldots,\overleftarrow{F}_p(U^p))$ alors on est ramené au cas précédent pour $\tilde{\eta}$ pour la mesure uniforme produit sur le cube unité.
On obtient la décomposition de Hoeffding de η à partir de celle de $\tilde{\eta}$ en remplaçant les U^i par $F_i(X^i)$. Cette méthode sera détaillée dans la partie 4.3.2.

2.3.3 Calcul pratique des indices de Sobol

Afin de calculer les indices de Sobol, il est classique d'utiliser des bases pour décomposer $L^2(\mu)$ en sous espaces orthogonaux simples lorsque μ est une mesure sur $\mathbb{R}^p$. Il en est ainsi des bases de Fourier, d'ondelettes, de celles construites sur des produits de polynômes orthogonaux dans le cas où μ est une mesure produit. L'espace $L^2(\mu)$ des fonctions $h(X^1,\ldots,X^n)$ de carré intégrable est considéré alors comme la limite croissante de sous espaces plus simples souvent de dimension finie. Par exemple pour le système exponentiel $\{\exp(i<n,x>), n\in\mathbb{Z}^p, x\in\mathbb{R}\}$, on peut considérer comme sous espaces approchant $L^2(\mu)$, les sous espaces des polynômes trigonométriques de degrés inférieurs ou égaux à d, le degré étant défini par $|n| = |n_1|+\cdots+|n_p|$ pour l'exponentielle $\exp(i<n,x>)$ et par $sup(|n|)$ pour une somme d'exponentielles. Ces méthodes sont classiquement utilisées pour approximer $\mathbf{E}(\eta(X^1,\ldots,X^p)|X^{i_1},\ldots,X^{i_q})$ car le calcul des espérances conditionnelles amène à des quadratures compliquées si η est compliquée, le type de calcul est :

$$\int_{\mathbb{R}^{p-q}} \eta(x^1,\ldots,x^p)\mathrm{d}\tilde{\mu}(x^{i'_1},\ldots,x^{i'_{p-q}})$$

où $\tilde{\mu}$ est la loi de $X^{i'_1},\ldots,X^{i'_{p-q}}$ avec $\{i'_1,\ldots,i'_{p-q}\} = \{1,\ldots,p\} - \{i_1,\ldots,i_q\}$.
Si $\mathrm{d}\mu$ a une densité μ alors l'espérance conditionnelle est le quotient de :
$\int_{\mathbb{R}^{p-q}} \eta(x^1,\ldots,x^p)\mu(x^1,\ldots,x^p)\mathrm{d}x^{i'_1}\ldots\mathrm{d}x^{i'_{p-q}}$ par $\int_{\mathbb{R}^{p-q}} \mu(x^{i'_1},\ldots,x^{i'_{p-q}})\mathrm{d}x^{i'_1}\ldots\mathrm{d}x^{i'_{p-q}}$.
Ces méthodes de calcul évitant la décomposition de Hoeffding interviennent dans la plupart des champs d'application de la théorie des probabilités et de la statistique, par exemple en théorie du signal ou de l'image. L'aspect technique du calcul est alors prépondérant. Détaillons le cas

des chaos polynomiaux largement utilisé en analyse de sensibilité comme le sont les méthodes de Fourier.

Polynômes de Chaos et analyse de sensibilité

Les polynômes orthogonaux $P_n(X)$ de $L^2(\nu)$, ν loi de probabilité sur $\mathbb{R}$, sont une famille de polynômes $P_m(X)$ (resp. P_n) de degrés m (resp. n) tels que :

$$\int_{\mathbb{R}} P_m(X)P_n(X)d\nu(X) = c_{mn}\delta_{mn} \quad m \neq n$$

On peut montrer qu'il existe une famille unique de polynômes orthogonaux et formant une base orthogonale complète pour l'espace $L^2(\nu)$. Les polynômes d'Hermites H_n, de Legendre (modifiés) L_n et de Laguerre $\mathbb{L}_n$ sont des exemples de familles de polynômes orthogonaux pour des lois gaussiennes $\mathcal{N}(0,1)$, uniformes $U(0,1)$ et exponentielles respectivement.

$$H = H_0 \oplus H_1 \oplus \cdots \oplus H_p$$

H_k est l'espace de Hilbert engendré par le produit de type :

$$P_{i_1}(X^{i_1}) \dots P_{i_k}(X^{i_k})$$

avec $(i_1, \dots, i_k) \in (1, \dots, p)^k$. Ces produits sont appelés dans le domaine de la sensibilité : polynôme de chaos.

Dans le cas de la loi uniforme et des polynômes de Legendre, on peut voir l'analogie avec la décomposition de Hoeffding. Les produits de polynômes remplacent et approchent la base de Hoeffding.

Nous avons donc les développements suivants, si
$a^{(1\dots d)}_{i_1\dots i_q} = \int \eta(X^1, \dots, X^p)P(X^{i_1}) \dots P(X^{i_q})\mathrm{d}\nu(X^{i_1}) \dots \mathrm{d}\nu(X^{i_q})$:

$$\begin{array}{lcl} \eta_0 & = & a_0^{(0)} \\ \eta_i(X^i) & = & \sum_{n\in\mathbb{N}} a_n^{(i)} L_n(X^i) \\ \eta_{ij}(X^i, X^j) & = & \sum_{n,m\in\mathbb{N}} a_{mn}^{(ij)} L_m(X^i)L_n(X^j), \forall j > i \\ \dots & & \\ \eta_{1\dots d}(X^1, \dots, X^p) & = & \sum_{i_1,\dots i_p\in\mathbb{N}} a_{i_1\dots i_p}^{(1\dots d)} L_{i_1}(X^1) \dots L_{i_s}(X^p), \end{array} \tag{I.19}$$

qui conduisent au développement sur le chaos polynomial suivant :

$$\eta(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha}\in\mathbb{N}^d} a_{\boldsymbol{\alpha}} L_{\boldsymbol{\alpha}}(X^{i_1}, \dots, X^{i_p}), \tag{I.20}$$

où on a posé ; $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p) \in \mathbb{N}^p$ et $L_{\boldsymbol{\alpha}} = L_{\alpha_1} \times L_{\alpha_2} \times \dots L_{\alpha_p}$ avec $L_0 = 1$. On note $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_p$.

La propriété d'orthonormalité, assure $\eta_0 = a_0$ ainsi que

$$D = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d / \{\mathbf{0}\}} a_{\boldsymbol{\alpha}}^2 - a_{\mathbf{0}}^2. \tag{I.21}$$

La variance partielle due à l'interaction entre $\{i_1, \ldots, i_s\}$ s'écrit,

$$D_{i_1 \ldots i_s} = \sum_{\boldsymbol{\alpha} \in \mathcal{I}_{i_1,\ldots,i_s}} a_{\boldsymbol{\alpha}}^2 - a_{\mathbf{0}}^2. \tag{I.22}$$

où, $\mathcal{I}_{i_1,\ldots,i_s} = \left\{ \boldsymbol{\alpha} : \begin{array}{lll} \alpha_k > 0 & , & k \in (i_1, \ldots, i_s) \\ \alpha_k = 0 & , & k \notin (i_1, \ldots, i_s) \end{array} \right\}$

Remarque 3. *Cette décomposition n'est pas la seule. On utilise aussi la décomposition en polynômes de degré homogène. Par exemple pour les chaos Gaussiens (de Wiener)*

$$\begin{aligned} H_0^1 &= Constante \\ H_1^1 &= \{X^i, i = 1; \ldots, p\} \\ H_2^1 &= \{X^i X^j, (X^i)^2 - 1; i = 1, \ldots, p\} \\ &\ldots \end{aligned}$$

On remarque que la connaissance des coefficients $a_{\boldsymbol{\alpha}}$ permet de caractériser complètement l'incertitude sur Y. En pratique, le développement est tronqué jusqu'à un certain ordre polynomial M de sorte que l'on travaille avec un meta-modèle approché :

$$\eta(\boldsymbol{X}) \simeq \hat{\eta}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d}^{|\boldsymbol{\alpha}| \leq M} a_{\boldsymbol{\alpha}} L_{\boldsymbol{\alpha}}(X^{i_1}, \ldots, X^{i_s}), \tag{I.23}$$

conduisant aux estimateurs suivants :

$$\begin{aligned} \hat{D} &= \sum_{\boldsymbol{\alpha} \in \mathbb{N}^d / \{\mathbf{0}\}}^{|\boldsymbol{\alpha}| \leq M} a_{\boldsymbol{\alpha}}^2 - a_{\mathbf{0}}^2, \\ \hat{D}_{i_1 \ldots i_s} &= \sum_{\boldsymbol{\alpha} \in \mathcal{I}_{i_1,\ldots,i_s}}^{|\boldsymbol{\alpha}| \leq M} a_{\boldsymbol{\alpha}}^2 - a_{\mathbf{0}}^2, \\ \hat{S}_{i_1 \ldots i_s} &= \frac{\hat{D}_{i_1 \ldots i_s}}{\hat{D}}. \end{aligned} \tag{I.24}$$

Le nombre de coefficients dans Eq. (I.23) à déterminer est égal à : $n = \dfrac{(M+d)!}{M!d!}$, ce qui, par une méthode non-intrusive, nécessite un nombre de simulations $N \geq \dfrac{(M+d)!}{M!d!}$, renvoyant au problème du "fléau de la dimension". Il existe différentes techniques pour construire des polynômes de chaos correspondant à η donné en évitant des quadratures qui étaient le handicap principal du développement de Hoeffding.

Calcul des coefficients des polynômes de chaos par régression

La détermination des coefficients du PC par régression consiste, à partir d'un échantillon $\boldsymbol{X}$ de taille $N \geq \dfrac{(M+d)!}{M!d!}$, à calculer le meilleur jeu de coefficients $a_{\boldsymbol{\alpha}}$ au sens des moindres carrés,

$$\hat{a}_{\boldsymbol{\alpha}} = (\underline{\underline{L}}^T \underline{\underline{L}})^{-1} \underline{\underline{L}}^T \underline{y} \tag{I.25}$$

où, $\underline{\underline{L}}$ est une matrice $N \times n$ d'éléments $L_{ij} = L_i(\boldsymbol{x}^{(j)})$ avec $L_i \in L_{\boldsymbol{\alpha}}$ et $\underline{y}$ est un vecteur d'éléments $y_j = \eta(\boldsymbol{x}^{(j)})$.

Plusieurs stratégies d'échantillonnage peuvent être employées :

— *Régression probabiliste*
Cette technique d'échantillonnage, initialement proposé par Tatang et al. [117], consiste à n'affecter aux facteurs d'entrées que des valeurs correspondant aux $(M+1)$ racines du polynôme de degré $(M+1)$ (i.e. $L_{M+1}(X)$). Dans sa version originale, il fallait construire toutes les combinaisons possibles soit un nombre de simulations $N = d^{M+1}$. Par la suite, Huang et al. [63] et Sudret [114] ont proposé une stratégie permettant de ramener la taille de l'échantillon au nombre d'inconnues $N \simeq \dfrac{(M+d)!}{M!d!}$. L'idée consistant à choisir parmi les d^{M+1} combinaisons possibles les N points les plus proches de zéro et garantissant que $(\underline{\underline{L}}^T \underline{\underline{L}})$ est inversible.

— *Collocation stochastique*
Un échantillonnage de type Monte Carlo (LHS, QMC, ...) est alors employé. En pratique, des problèmes liés au sur-apprentissage peuvent survenir et l'emploi d'une technique de régularisation est nécessaire. Blatman et Sudret proposent dans [9] (voir aussi [10]) une stratégie d'échantillonnage (nested-LHS) afin de construire un PC creux pour un nombre minimal de simulations.

Calcul des coefficients des polynômes de chaos par projection

La détermination des coefficients du PC par projection consiste à exploiter la propriété d'orthonormalité des PC.

A partir d'un échantillon $\boldsymbol{x}$, les coefficients $a_{\boldsymbol{\alpha}}$ sont calculés de la façon suivante :

$$\hat{a}_{\boldsymbol{\alpha}} = \sum_{i=1}^{N} w_i L_{\boldsymbol{\alpha}}(\mathbf{X}^{(i)}) \eta(\mathbf{X}^{(\mathbf{i})}) \tag{I.26}$$

où les w_i sont des poids fonction de la stratégie d'échantillonnage adoptée.

— *Collocation stochastique.*
Identique au cas précédent et dans ce cas, les poids valent $w_i = 1$.

— *Smolyak quadrature.*
C'est une amélioration de la méthode "naturelle" de quadrature. Crestaux et al. [21] emploient cette technique d'échantillonnage afin de calculer les indices de Sobol. Elle est due à Smolyak [107] et est également très utilisée en construction de métamodèles par interpolation (cf. [74], [14]). Elle est plus économique que la quadrature de Gauss.

Remarque 4. *L'emploi des PC en tant que modèle simplifié (émulateur) est sujet à des conditions de régularité de la fonction η ([21], [9]).*

Les méthodes de Fourier : méthode FAST

Ce sont d'autres exemples des méthodes ANOVA. Elles peuvent être plus efficaces pour certains problèmes multidimensionnels mais présentent des difficultés de calcul, ainsi que des difficultés théoriques.

On se place encore dans le cas de variables aléatoires indépendantes et identiquement distribuées uniformes ($i.i.d$) et d'un modèle donné par :

$$Y = \eta(X^1, \ldots, X^p)$$

Y est maintenant considéré comme un signal échantillonné en N points :

$$\{x^{1,(k)}, \ldots, x^{p,(k)}, k = 1, \ldots, N\}$$

Pour chaque facteur X^i, on choisit une fréquence privilégiée ω_i.

L'ensemble $\{\omega_i, 1 \leq i \leq p\}$ est le spectre discret associé au problème.

On considère l'ensemble $\{s_k = \dfrac{2\pi(k-1)}{N}, 1 \leq k \leq N\}$ qui est la discrétisation du problème.

Pour chaque facteur i, on choisit le plan ou "courbe" d'échantillonnage :

$$\forall 1 \leq i \leq p,\ \forall 1 \leq k \leq N,\quad x^{i,(k)} = \frac{1}{\pi}\arcsin(\sin(\omega_i s_k + \phi_i)) + \frac{1}{2}$$

où les ϕ_i sont des variables aléatoires $i.i.d$, représentant la phase.

Le point de départ de la décomposition de la variance est l'orthogonalité des exponentielles $\exp(i) = \{\exp(k_1\omega_1 + \cdots + k_p\omega_p),\ (k_1, \ldots, k_p) \in \mathbb{Z}^p\}$

En fait, le choix des fréquences d'entrées, amène de par les discrétisations, à des phénomènes d'interférences que l'on peut interpréter comme des interactions.

Les coefficients de Fourier seront estimés et donnés sous la forme :

$$\hat{c}_{k_1\omega_1+\cdots+k_p\omega_p} = \frac{1}{N}\sum_{j=1}^{N} \eta(x^{1,(j)}, \ldots, x^{p,(j)}) e^{-2i\pi(j-1)(k_1\omega_1+\cdots+k_p\omega_p)/N} \tag{I.27}$$

donc par une somme discrète, approchant l'intégrale, calcul fondé sur des propriétés d'équidistribution des $\boldsymbol{x}^{(k)}$.

A partir de là, l'égalité de Parseval permet d'estimer les composantes de la variance. Ainsi un estimateur de la variance est :

$$\hat{V}(Y) = \frac{1}{N}\sum_{k=1}^{N/2} c_k^2$$

avec $c_k^2 = \frac{1}{N}\sum_{k=1}^{N} h(x^{1,(k)},\ldots,x^{p,(k)})e^{-2i\pi\frac{(k-1)}{N}}$

$\hat{V}_i$ estimateur de $\mathbf{Var}(Y|X^i)$ est estimé par :

$$\hat{V}_i = \frac{1}{N}\sum_{k=1}^{N_h} c^2_{k\omega_i}$$

où N_h est un coefficient de seuillage (interprété comme la première harmonique négligeable)

$$\hat{V}_{ij} = \frac{1}{N}\sum_{1\leq k+l\leq N_{kl}} c^2_{k\omega_i+l\omega_j}$$

N_{kl} s'interprète de la même manière que N_h. Il représente l'estimateur de $\mathbf{Var}(Y|X^iX^j) - \mathbf{Var}(Y|X^i) - \mathbf{Var}(Y|X^j)$ si les phénomènes d'interférence sont négligeables. Ces phénomènes surviennent aux différents ordres, par exemple à l'ordre 2 si il existe des entiers (a,b) tels que :

$$a\omega_1 + b\omega_2 = 0$$

Historiquement la méthode FAST a été introduite dans les années 70 par Cukier et al. au travers d'une série de quatre articles [22], [104], [24] et [23].

Le choix des fréquences est très problématique avec la méthode FAST classique. La méthode RBD-FAST (Random Balance Design) (Tarantola et al. [116]) permet de s'affranchir de la plupart des inconvénients de la méthode FAST classique en particulier la dépendance du nombre de simulations à la dimension du modèle (critère de Shannon) et le choix des fréquences.

Saltelli et al. [100] a étendu la méthode FAST classique au calcul des indices de sensibilité totaux permettant de déterminer tous les $(S^{X^i}, S^{X^i}_{tot})$ en $d \times N$ simulations tandis que Mara [83] l'a fait pour RBD-FAST. En outre, dans ce dernier article l'auteur propose également des méthodes pour calculer les indices des sensibilité d'ordre 2, 3, ...

Plischke [94] a étendu la méthode RBD-FAST au cas de l'échantillonnage non périodique et propose un correcteur de biais inhérent dans la version originale de la méthode RBD-FAST.

2.3.4 Méthode d'estimation basée sur l'échantillonnage : Méthode Pick and Freeze

Si l'on a un modèle $Y = \eta(\boldsymbol{U}, \boldsymbol{V})$ où $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{V})$ est l'entrée et Y la sortie, dans le cas où $\boldsymbol{U}$ et $\boldsymbol{V}$ sont indépendants et la relation η est connue on peut utiliser pour estimer la sensibilité des décompositions du type Hoeffding ou l'ensemble des méthodes vues précédemment. On peut aussi utiliser la méthode Pick and Freeze. Cette approche est celle initialement proposée par Sobol dans [108] pour estimer les indices de sensibilité. En effet ces derniers reposent sur les moments d'ordre 1 et 2 et peuvent être calculés à l'aide de simulations de type (Quasi) Monte Carlo ([102],[79]) de $\boldsymbol{X}^{(i)}$ sous la forme $(X^{1,(i)},\ldots,X^{p,(i)})_{i=1,\ldots,N}$ et la mesure des sorties correspondantes $(Y^{(i)})_{i=1,\ldots,N}$. La fonction η joue dans ce cadre le rôle d'une boîte noire

permettant de simuler la relation "entrée-sortie" sans que sa description mathématique ne soit utilisée.

Posons $\boldsymbol{X} = (\boldsymbol{U}, \boldsymbol{V})$ où $\boldsymbol{U} = \{X^{i_1}, \ldots, X^{i_k}\}$ regroupe k facteurs d'entrée et $\boldsymbol{V}$ le complément ($d-k$ facteurs), $Y = \eta(\boldsymbol{U}, \boldsymbol{V})$

La méthode Pick and Freeze repose sur le lemme de Sobol suivant :

Lemme 1. *Si $\boldsymbol{U}$ et $\boldsymbol{V}$ sont indépendantes alors* $\mathbf{Var}(\mathbf{E}(Y|\boldsymbol{U})) = \mathbf{Cov}(Y, Y^U)$ *où* $Y^U = \eta(\boldsymbol{U}, \boldsymbol{V}')$, $\boldsymbol{V}'$ *copie indépendante de* $\boldsymbol{V}$.

En effet, en remarquant que :

$$\begin{aligned}\mathbf{Var}(\mathbf{E}(Y|\boldsymbol{U})) &= \mathbf{E}(\mathbf{E}(Y|\boldsymbol{U})^2) - \mathbf{E}(\mathbf{E}(Y|\boldsymbol{U}))^2 \\ &= \mathbf{E}(\mathbf{E}(Y|\boldsymbol{U})^2) - \mathbf{E}(Y)^2 \\ \mathbf{Cov}(Y, Y^U) &= \mathbf{E}((YY^U)) - \mathbf{E}(YY^U)\end{aligned}$$

et Y, Y^U étant indépendants conditionnellement à $\boldsymbol{U}$ et de même loi

$$\begin{aligned}&\mathbf{E}(YY^U|\boldsymbol{U}) = \mathbf{E}(Y|U)\mathbf{E}(Y^U|\boldsymbol{U}) = \mathbf{E}(Y|\boldsymbol{U})^2 \\ &\mathbf{E}(\mathbf{E}(YY^U|\boldsymbol{U})) = \mathbf{E}(YY^U)\end{aligned}$$

Alors on peut réécrire l'indice de Sobol de la façon suivante :

$$S^U = \frac{\mathbf{Cov}(Y^U, Y)}{\mathbf{Var}(Y)} \tag{I.28}$$

L'estimateur Pick and Freeze consiste à prendre un estimateur empirique du numérateur et du dénominateur. Cette méthode est robuste ([69]) et sa vitesse de convergence est indépendante du nombre de variables d'entrée.

On choisit un échantillon de taille N, $\{(Y^{(1)}, Y^{U,(1)}), \ldots (Y^{(N)}, Y^{U,(N)})\}$. Un estimateur naturel est :

$$\widehat{S}_N^U = \frac{\frac{1}{N}\sum_{i=1}^N Y^{(i)}Y^{U,(i)} - (\frac{1}{N}\sum_{i=1}^N Y^{(i)})(\frac{1}{N}\sum_{i=1}^N Y^{U,(i)})}{\frac{1}{N}\sum_{i=1}^N (Y^{(i)})^2 - (\frac{1}{N}\sum_{i=1}^N Y^{(i)})^2} \tag{I.29}$$

Janon et al. ([69]) proposent un autre estimateur optimal en termes de variance asymptotique :

$$\widehat{S}_N^U = \frac{\frac{1}{N}\sum_{i=1}^N Y^{(i)}Y^{U,(i)} - (\frac{1}{2N}\sum_{i=1}^N Y^{(i)} + Y^{U,(i)})^2}{\frac{1}{N}\sum_{i=1}^N \frac{(Y^{(i)})^2 + (Y^{U,(i)})^2}{2} - (\frac{1}{N}\sum_{i=1}^N \frac{Y^{(i)} + Y^{U,(i)}}{2})^2} \tag{I.30}$$

La qualité de l'estimateur repose sur l'échantillonneur utilisé. Sobol préconise l'utilisation des échantillonneurs à faibles discrépances aussi appelés échantillonneurs quasi-Monte Carlo. L'un des plus performants repose sur les séquences LP_τ qu'il a lui même développé. En outre, pour garantir la qualité de l'estimateur il est recommandé d'utiliser la technique du *Bootstrap* [37] qui fournit un intervalle de confiance aux grandeurs estimées [103]. Saltelli [98] propose une stratégie économique afin de déterminer tous les effets principaux et totaux ; stratégie basée sur la technique RBD (acronyme pour Random Balance Design).

Propriétés asymptotiques :

Dans [69], Janon et al. montrent que dans ce cas, les estimateurs $\widehat{S}_N^U$ convergent vers S^U quand $N \to \infty$ et de plus si $\mathbf{E}|Y|^4 < \infty$, alors :

$$\sqrt{N}(\widehat{S}_N^U - S^U) \to N(0, \Gamma) \tag{I.31}$$

la valeur de Γ étant calculable en fonction des moments d'ordre inférieur ou égal à 4. Il s'agit là d'un résultat asymptotique qui permet d'obtenir des intervalles de confiances approchés.

De plus l'estimateur est asymptotiquement efficace c'est-à-dire que γ est minimal. Enfin cet estimateur possède différentes propriétés de robustesse, notamment si le couple entrée-sortie est bruité [48] [47], [70].

Chapitre 3

Modèles statistiques couramment utilisés

3.1 Régressions linéaires et non linéaires

Modèles linéaires

Les modèles de régressions linéaires sont les modèles les plus largement utilisés pour effectuer une analyse de sensibilité dans le contexte de l'analyse énergétique des bâtiments. Ces modèles présentent deux avantages ; ils sont rapides à mettre en œuvre et faciles à interpréter.

Considérons le modèle linéaire simple suivant :

$$Y = \theta X + \varepsilon$$

Remarquons que c'est un modèle à sortie bruitée. Si du point de vue de la sensibilité ce bruit ne joue pas de rôle important, il est néanmoins vrai que la plupart des méta-modèles statistiques d'entrée sont bruités et que physiquement cela peut être aussi le cas des sorties.

X : la matrice des N observations du vecteur $\boldsymbol{X}$ de dimension p

Y : le vecteur de sortie des N observation correspondant à la matrice X.

θ : le vecteur des paramètres inconnus de dimension p.

ε : est centré de variance constante.

On suppose, de plus, ici que $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ et que ses différentes réalisations $\varepsilon^{(1)}, \ldots, \varepsilon^{(N)}$ sont indépendantes.

Remarque 5. *De nombreux résultats sont vrais sans qu'il y ait besoin de l'hypothèse de normalité sur le bruit ε. Cependant cette hypothèse devient essentielle si on veut utiliser la théorie du maximum de vraisemblance et déterminer la loi des statistiques de test.*

Une fois le modèle choisi, il faut estimer les paramètres. Une première manière d'aborder l'estimation des composantes du vecteur θ est de minimiser la norme du vecteur résiduel $\varepsilon = Y - \theta X$ soit :

$$\hat{\theta}_N = \min_{\theta \in \Theta} \|Y - \theta X\|^2 \tag{I.1}$$

Alors l'estimateur classique des moindres carrés est :

$$\hat{\theta}_N = (X^*X)^{-1}X^*Y$$

avec :

- $\hat{\theta}_N \sim \mathcal{N}(\theta, \sigma^2(X^*X)^{-1})$
- $\frac{(N-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{N-p}$ avec $\hat{\sigma}^2 = \frac{\|Y - \hat{\theta}_N X\|^2}{(N-p)}$
- $\hat{\theta}_N$ et $\hat{\sigma}^2$ sont indépendants

On peut mesurer l'adéquation du modèle aux données via le rapport :

$$0 \leq R^2 = \frac{SSR_{eg}}{SST} \leq 1$$

où $SST = \|\hat{Y} - \overline{Y}\|^2$ et $SSR_{eg} = \|Y - \overline{Y}\|^2 = \mathbf{Var}(Y)$
(SST : Sum of Squares Total et SSR : Sum of squares Regression)
avec $\overline{Y}$ le vecteur de la moyenne de Y et $\hat{Y}$ la prédiction de Y , $\hat{Y} = \hat{\theta}_N X$

Les régions de confiance sont définies par l'ensemble des valeurs du vecteur des paramètres θ qui ne sont pas rejetées, au seuil donné, par le test du rapport des vraisemblances maximales. Cependant ce sont les intervalles de confiance qui apparaissent le plus souvent dans la littérature et les logiciels [3].
Les intervalles de confiance ne sont pas bien adaptés lorsqu'on veut considérer plusieurs paramètres simultanément, car ils ne tiennent pas compte de la dépendance des paramètres. Lorsque les paramètres de θ ne sont pas fortement corrélés, alors les régions définies par les intervalles de confiance (parallélépipèdes) sont de bonnes approximations des ellipsoïdes de confiance. Ce n'est plus le cas lorsque les paramètres sont fortement corrélés.

Remarque 6. *Les résultats de l'ajustement par moindres carrés du modèle linéaire général à un ensemble d'observations, peuvent être sensiblement modifiés par la suppression où la perturbation de certaines données. La représentation graphique des résidus permet de déceler des données aberrantes mais aussi de s'assurer de la validité du modèle.*

Dans le cas d'un modèle linéaire gaussien, l'espérance conditionnelle de Y par rapport à X^i (lorsque les variables sont centrées) vaut :

$$\mathbf{E}(Y|X^i) = \frac{\mathbf{Cov}(Y, X^i)}{\mathbf{Var}(X^i)} X^i$$

Pour hiérarchiser l'influence de toutes les entrées sur la sortie l'indice de sensibilité n'est autre que ρ le coefficient de corrélation linéaire :

$$\rho(X^j, Y) = \frac{\mathbf{Cov}(X^j, Y)}{\sqrt{\mathbf{Var}(X^j)}\sqrt{\mathbf{Var}(Y)}}$$

Un calcul élémentaire montre que la sensibilité est donnée par $S^{X^j} = \rho(X^j, Y)$. Si les entrées sont indépendantes alors $\mathbf{Cov}(X^j, Y) = \theta_j \mathbf{Var}(X^j)$.

Régression non linéaire

On considère le modèle non linéaire suivant :

$$Y = \eta(\theta, \boldsymbol{X}) + \varepsilon$$

$\boldsymbol{X}$ le vecteur d'entrée de dimension p

ε : bruit $i.i.d$ et de variance σ^2

$\boldsymbol{X}^{(i)}$ est une suite $i = 1, \ldots, N$ de réalisations.

$\boldsymbol{Y}^{(1;N)} = \{Y^{(i)}, i = 1, \ldots, N\}$.

Estimateur des moindres carrés de θ : On note :

$$\hat{\theta}_N = \underset{\theta \in \Theta}{argmin} \frac{1}{N} \sum_{i=1}^{N} (Y^{(i)} - \eta(\theta, \boldsymbol{X}^{(i)})^2 \tag{I.2}$$

un estimateur des moindres carrés de θ,

On suppose Θ compact avec $\theta_0 \in \overset{\circ}{\Theta}$.

Proposition 2. *Tout estimateur des moindres carrés de θ_0 appartenant à l'intérieur de Θ est solution des équations normales :*

$$(Y - \eta(\boldsymbol{X}, \theta))^* \nabla \eta(\boldsymbol{X}, \theta) = 0$$

où D est le gradient

Régions de confiance : cas Gaussien

Théorème 1. *Une région de confiance pour le paramètre θ au niveau approché $1 - \alpha$ est définie par les valeurs de θ pour lesquelles si $C_N(\theta) = \frac{1}{N} \sum_{i=1}^{N} (Y - \eta(\theta, \boldsymbol{X}^{(i)}))^2$:*

Si σ^2 est connu :

$$C_N(\theta) - C_N(\hat{\theta}) \leq \frac{\sigma^2}{N} \chi_p^2 (1 - \alpha)$$

où $\chi_p^2 (1 - \alpha)$ est le quantile d'ordre $1 - \alpha$ de loi du χ^2 à p degrés de liberté.

Si σ^2 inconnu :

$$\frac{C_N(\theta) - C_N(\hat{\theta})}{C_N(\hat{\theta})} \leq \frac{p}{N - p} F_{p,N-p} (1 - \alpha)$$

où $F_{p,N-p} (1 - \alpha)$ est le quantile d'ordre $1 - \alpha$ de loi de Fisher F à p et $N - p$ degrés de liberté.

Extensions :

Les modèles de régression présents précédemment possèdent des hypothèses trop restrictives (erreurs centrées et intervenant de manière additive sur la réponse moyenne, homoscédasticité (même variance),...) pour couvrir l'ensemble des champs d'application usuels relevant d'un modèle de régression. Nelder et Wedderburn ont introduit la notion de modèle de régression généralisé (Y suit une loi appartenant à une structure exponentielle).

Bibliographie : [99], [7], [2], [101]

3.2 Modèles additifs généralisés

Ces méta-modèles statistiques sont très utiles dans une phase exploratoire avant un choix définitif. On peut dans le cas d'entrées indépendantes, les considérer comme une approximation de la représentation de Hoeffding, le plus souvent bruitée en Y. Le modèle le plus simple est :

$$Y = \sum_{j=1}^{p} h_j(X^j) + \varepsilon$$

où les h_j sont des fonctions inconnues que l'on doit estimer à partir d'un échantillon $(Y^{(1;N)}, \boldsymbol{X}^{(1;N)}) = (Y^{(i)}, \boldsymbol{X}^{(i)})_{i=1,\dots,N}$, $\boldsymbol{X}^{(i)} = (X^{1,(i)}, \dots, X^{p,(i)})$.

Une des méthodes d'estimation les plus utilisées est l'estimation par splines cubiques particulièrement dans le cas ou les valeurs de $\boldsymbol{X}$ sont bornées. Dans ce cadre la seule hypothèse portant sur les fonctions inconnues h_j est qu'elles sont de classe $\mathcal{C}^2$. Le principe d'estimation est alors de minimiser en $(h_1, \dots, h_p)$ le critère :

$$S(h_1, \dots, h_p) = \sum_{i=1}^{N} |Y^{(i)} - \sum_{j=1}^{p} h_j(X^{j,(i)}|^2 + \sum_{j=1}^{p} \lambda_j \int |h_j''(t)|^2 \mathrm{d}t \tag{I.3}$$

Cette méthode est simple à mettre en œuvre car la minimisation se ramène à un problème linéaire de dimension N.

Si $\{N_q(x), q = 1, \dots, m\}$ est une base de splines dans le cas $p = 1$, le critère à minimiser se réduit en posant $h(x) = \sum_{q=1}^{N} \theta_q N_q(x) = N(x)\theta$ à :

$$(Y - N\theta)^*(Y - N\theta) + \lambda\theta^*\Omega_N\theta$$

où $\Omega_N = \{\int N_j''(t)N_k''(t)\mathrm{d}t, 1 \leq i, j \leq m\}$. On en déduit :

$$\hat{\theta} = (N^*N + \lambda\Omega_N)N^*Y$$

On obtient alors $\hat{h}$ estimateur de h.
Le problème principal est de choisir λ. Ceci peut se faire par validation croisée [57]. Dans le cas $p > 1$, une méthode analogue et récursive utilisant les splines cubiques peut être utilisée.

Un exposé détaillé se trouve dans [57] et pour la partie plus algorithmique dans [58]. Cette technique peut s'étendre au cas où l'on souhaite estimer les termes d'ordre 2 représentant les interactions par paires. Le modèle est alors :

$$Y = \sum_{j=1}^{p} h_j(X^j) + \sum_{1 \leq j \neq k \leq p} h_{j,k}(X^j, X^k) + \varepsilon$$

Les bases de splines cubiques matricielles sont simplement la base obtenue par produit tensoriel et la pénalisation est du type :

$$\lambda \int \left(\sum_{i,k=1}^{p} |\frac{\partial^2 h_{j,k}}{\partial x^i \partial x^k}|^2 + |\frac{\partial^2 h_{j,k}}{(\partial x^i)^2}|^2 + |\frac{\partial^2 h_{j,k}}{(\partial x^k)^2}|^2 \right) \mathrm{d}t$$

En ce qui concerne les calculs préliminaires de sensibilité à partir des modèles additifs d'ordre 1 dans le cas où les entrées sont indépendantes, on remarque que l'espérance conditionnelle se calcule simplement par :

$$\mathbf{E}(Y|X^j) = h_j(X^j) + constante$$

La constante n'affecte pas la variance et donc l'indice de sensibilité est :

$$S^{X^j} = \frac{\mathbf{Var}(h_j(X^j))}{\sum_{j=1}^{p} \mathbf{Var}(h_j(X^j))} \tag{I.4}$$

Ce calcul peut se faire assez aisément par simulation.
Dans le cas où les variables ne sont pas indépendantes, nous verrons ultérieurement ce qu'il est possible de faire pour calculer cette sensibilité.

3.3 Champs gaussiens et krigeage

Il s'agit de méta-modèles ayant des similitudes mais aussi de profondes différences avec les modèles de régressions. La première différence est l'indexation des variables d'entrées et de sorties.

Le modèle suivant :

$$Y(\boldsymbol{x}) = m(\boldsymbol{x}) + Z(\boldsymbol{x}) \tag{I.5}$$

où

- $m : \boldsymbol{x} \in \mathbb{R}^p \to m(\boldsymbol{x}) \in \mathbb{R}$ est une fonction moyenne (appelée aussi tendance). C'est la partie déterministe du modèle.
- $\boldsymbol{x} \mapsto Z(\boldsymbol{x})$ est un champ gaussien centré défini par :

$$\mathbf{E}(Z(\boldsymbol{x})) = 0 \tag{I.6}$$
$$\mathbf{Cov}(Z(x^i), Z(x^j)) = k(x^i, x^j) = \sigma^2 R(x^i - x^j) \tag{I.7}$$

où $\sigma^2 \in \mathbb{R}$ désigne la variance de Z et R est sa fonction de corrélation.

La partie déterministe se limite le plus souvent à l'utilisation de polynômes de degré 0 ou 1 :

$$m(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j = h(\boldsymbol{x})\beta \tag{I.8}$$

$\beta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^p$ sont les paramètres de régression et $h(\boldsymbol{x}) = (1, x^1, \ldots, x^p)$ est le vecteur de régression au point $\boldsymbol{x}$. Cette formulation se généralise aisément à d'autre bases de fonctions de régression.

Cependant de nombreux auteurs conseillent de ne considérer qu'une fonction constante pour la tendance ($m(\boldsymbol{x}) = \beta_0$) arguant que le processus gaussien $Z(\boldsymbol{x})$ est suffisant pour capter les non linéarités et interactions du modèle. A contrario d'autres auteurs estiment que cette fonction tendance est importante pour limiter le rôle de la partie stochastique du modèle gaussien à la modélistation des fluctuations rapides du modèle. Enfin la tendance déterministe offre une opportunité d'introduire d'éventuelles informations a priori inhérente à la physique du phénomène simulé.

Remarque 7. *Le choix de la fonction R est une des difficultés principales pour l'usage de ce type de méta-modèle. R est rarement estimée et les problèmes d'anisotropie sont nombreux.*

L'estimation de m ou h dépend de la structure de covariance k supposée connue. On utilise alors une méthode voisine de celle utilisée précédemment pour les modèles additifs non linéaires.

La fonction de covariance $k(.,.)$ est aussi appelée noyau reproduisant et l'espace des combinaisons linéaires $\sum_{i=1}^{n} \lambda_i k(\boldsymbol{x}^{(i)}, \cdot)$, munies du produit scalaire défini par :

$$< k(\boldsymbol{x}^{(i)}, \cdot); k(\boldsymbol{x}^{(j)}, \cdot) >= k(\boldsymbol{x}^{(i)}; \boldsymbol{x}^{(i)})$$

peut être complété pour donner l'espace de Hilbert $\mathcal{H}_k$

Théorème 2. *Soit $(\boldsymbol{x}^{(i)}, y^{(i)})_{i=1,\ldots,N}$, l'ensemble d'apprentissage.*
Alors pour $\gamma > 0$ donné :

$$\underset{h \in \mathcal{H}}{argmin} \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - h(\boldsymbol{x}^{(i)}))^2 + \gamma \|h\|^2_{\mathcal{H}}$$

est unique et s'écrit sous la forme :

$$h(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}^{(i)}, \boldsymbol{x})$$

où $\alpha = (\alpha_1, \ldots, \alpha_N)$ est solution de l'équation :

$$(N\gamma I_N + K)\alpha = Y^N$$

où K matrice de taille $N \times N$ des $(k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}))_{1 \leq i,j \leq N}$ et $\boldsymbol{Y}^{(1;N)} = (Y^{(1)}, \ldots, Y^{(N)})$

On remarque que $h(\cdot)$ est évaluée grâce au noyau $k(\cdot;\cdot)$. Le choix de $k(\cdot;\cdot)$ est donc une étape importante.

Le krigeage est d'abord une manière d'extrapoler et de prédire la valeur de $Y(\boldsymbol{x})$ pour toute valeur $\boldsymbol{x}$.

On sait que la meilleure prédiction $\hat{Y}(\boldsymbol{x})$ de $Y(\boldsymbol{x})$ est :

$$\hat{Y}(\boldsymbol{x}) = \mathbf{E}(Y|Y^{(1)}, \ldots, Y^{(N)}) \tag{I.9}$$

avec $Y^{(i)} = Y(\boldsymbol{x}^{(i)})$. $\hat{Y}$ est de la forme $\sum_{i=1}^{N} \lambda_i Y(\boldsymbol{x}^{(i)})$. Le calcul de λ est simple.

Si $\Gamma = \left(k(x^{(i)}, x^{(j)})\right)_{1\leq i,j\leq N}$ alors :

$$\begin{aligned}\lambda(\boldsymbol{x}) &= \Gamma^{-1}\mathbf{Cov}(Y(\boldsymbol{x}), Y^{(N)}) \text{ avec } \boldsymbol{Y}^{(1;N)} = (Y^{(1)}, \ldots, Y^{(N)}) \\ &= \Gamma^{-1}(k(\boldsymbol{x}, \boldsymbol{x}^{(i)}))_{i=1,\ldots,N}\end{aligned}$$

La simulation conditionnelle du champ $Z(\boldsymbol{x})$ quand $(Z(\boldsymbol{x}^{(1)}), \ldots, Z(\boldsymbol{x}^{(N)}))$ est fixé et donc celle du champ $\sum_{i=1}^{N} \lambda_i(\boldsymbol{x}) Z(\boldsymbol{x}^{(i)})$ et donc de covariance $\sum_{1\leq i,j\leq N} \lambda_i(\boldsymbol{x})\lambda_j(\boldsymbol{x}')k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$. Des méthodes voisines sont utilisées dans le cas de processus temporels à temps discret ou continu. La nature du domaine des valeurs possibles de $\boldsymbol{x}$ est surtout importante pour le choix de la covariance de Z.

L'application du krigeage aux problèmes de sensibilité est développée dans Iooss et al. [66].

Chapitre 4

Analyse de sensibilité pour des entrées dépendantes

Les modèles du bâtiment possèdent des entrées corrélées la plupart du temps. Les paramètres du bâtiment peuvent être corrélés (la résistance thermique et l'inertie thermique sont liés par l'épaisseur du mur) aussi bien que les entrées dynamiques (la température d'une pièce dépend de la température de celles qui lui sont adjacentes).

Par conséquent, il est nécessaire de prendre en compte les effets de ces entrées corrélées dans l'application de l'analyse de sensibilité. Cependant, la plupart des recherches antérieures ne le font pas. La raison principale est que les méthodes disponibles sont relativement bien définies dans le cadre statique avec des paramètres indépendants.

Pour palier à ce problème, certains auteurs tels que Kucherenko et al. ([76]) proposent de garder la même définition de l'indice de Sobol du premier ordre mais présentent une méthode d'estimation qui ne nécessite pas l'indépendance des variables d'entrées.

D'autres comme Chastaing et al. ([17]) proposent une autre définition de l'indice en séparant l'indice en deux parties :

- $\mathbf{Var}(\eta_i(X^i)$ où η_i est le terme associé à i dans la formule de Hoeffding : qui mesure l'impact de la variable X^i sur la sortie
- $\Sigma_{i\notin v, v\neq\varnothing}\mathbf{Cov}(\eta_i(X^i), \eta_v(X^v))$: qui prend en compte la dépendance de X^i avec les autres entrées du modèle.

L'indice du premier ordre se réécrit alors de la manière suivante :

$$S^{X_i} = \frac{\mathbf{Var}(\eta_i(X^i)) + \Sigma_{i\notin v, v\neq\varnothing}\mathbf{Cov}(\eta_i(X^i), \eta_v(X^v))}{\mathbf{Var}(Y)} \tag{I.1}$$

Cette définition permet de distinguer la variabilité propre due à X^i elle même et à ces variabilités "imbriquées". Cette définition est difficile à interpréter car l'indice ne se trouve plus entre 0 et 1. Cependant étant basée sur la décomposition de Hoeffdding la somme de ces indices reste égale à 1.

On peut remarquer aussi que si les entrées sont indépendantes entre elles, la partie portant sur la covariance devient nulle et l'on obtient la définition classique de l'indice.

Dans la plupart des problèmes concrets, le modèle ou le méta-modèle entrée-sortie exige de manipuler des variables d'entrée dépendantes. Ce problème fait l'objet d'une littérature assez importante et variée. Nous présentons des points importants de cette littérature en insistant sur les méthodes que nous proposons essentiellement celles qui ont deux qualités :

— s'étendre au cas dynamique avec des entrées décrites sous forme de séries temporelles
— utiliser pour calculer la sensibilité des méthodes de simulation du type Pick and Freeze.

Les méthodes comme nous le verrons sont toujours ou presque utilisées pour des méta-modèles statistiques, elles deviennent plus complexes dès qu'il faut utiliser des échantillons d'apprentissage pour estimer les lois de probabilité gouvernant le modèle.

Soulignons aussi quelques points. Indépendance et orthogonalité peuvent être du point de vue pratique très différentes. La corrélation ne traduit pas toujours convenablement la dépendance et en ce qui concerne les calculs de sensibilité, il faut être prudent. La dépendance fait perdre a priori deux outils très importants :

— la formule de Hoeffding (sous ses différentes formes) permettant des développements dans des bases diverses de l'espace L^2, type Fourrier, polynômes de chaos, ..., qui permet d'approximer la sortie afin de limiter la complexité des calculs lors de l'évaluation des indices.
— le lemme de Sobol sur lequel repose la méthode Pick and Freeze ne vaut que pour des entrées indépendantes.

Nous examinerons d'abord des méthodes encore en cours de développement mais qui semblent les plus naturelles. Ce sont celles fondées sur une formule de Hoeffding pour des variables dépendantes. Nous étudierons ensuite les méta-modèles fondés sur la notion de copule. Nous regarderons quelques cas en rapport avec la méthode que nous développons au paragraphe suivant, en particulier celui des copules Gaussiennes et accessoirement elliptiques.

Nous examinerons les méthodes fondées sur des transformations de loi pour se ramener à des variables indépendantes et uniformes ou Gaussiennes. Ces développements sont tous anciens mais nous les adaptons et les utilisons pour obtenir des formules de Hoeffding en variables dépendantes grâce auxquelles nous appliquons la méthode Pick and Freeze. Dans la dernière partie nous regardons les problèmes liés à la non connaissance des lois du méta-modèle et à la nécessité d'un apprentissage. Cela limite la valeur des modèles de copules et s'étend plus ou moins bien aux méthodes de transformation.

4.1 Formule de Hoeffding en variables dépendantes et modèles hiérarchiques

Le point de départ est l'article de Stone [113] très utilisé en statistique. Ce travail a été étendu en vue des études de sensibilité par Hooker [61] et prolongé notamment par le travail de Chastaing et al. [18], [17].

Soit $\mathcal{S}$ l'ensemble des parties non vides de $(1, \ldots, p)$, p la dimension du vecteur d'entrée $\boldsymbol{X}$.

Soit $\boldsymbol{X}^u = \{X^l, l \in u, u \in \mathcal{S}\}$ et H^u, pour $u \in \mathcal{S}$, l'espace de Hilbert des fonctions de carré intégrable $\boldsymbol{X}^u$-mesurables.
On note $H_u^0 = \{h \in H^u, \forall k \in H^v, \forall v \subset u, < h, k >= 0\}$

$H^0_\varnothing$ sera l'espace des constantes et $H^0 = \{h(\boldsymbol{X}) = \sum_{u\in\mathcal{S}\cup\varnothing} h_u(\boldsymbol{X}^u), h_u \in H^0_u\}$.

Les éléments de H^0 seront donc ceux qui admettent une décomposition que nous appelons hiérarchique (pour l'ordre de l'inclusion).

Faisons d'abord une remarque sur l'indice de Sobol dans le cas indépendant. Soit $S^{\boldsymbol{X}^u}$ l'indice associé à u donné par : $S^{\boldsymbol{X}^u} = \dfrac{\mathbf{Var}(\eta_u)}{\mathbf{Var}(Y)}$ où η_u est le terme associé à u dans la formule de Hoeffding. On sait que :

$$\mathbf{Var}(\eta_u) = \mathbf{Var}\left(\mathbf{E}(Y|\boldsymbol{X}^u)\right) - \sum_{v\subset u}(-1)^{-u-v}\mathbf{Var}\left(\mathbf{E}(Y|\boldsymbol{X}^v)\right)$$

Dans le cas dépendant, ce type de résultat va être étendu sous des conditions assez contraignantes mais dont la nécessité n'est pas évidente et qui ne sont pas intuitives. Outre des conditions de domination de la loi de $\boldsymbol{X}$ par des mesures produits, la condition suffisante est :

$$\exists M,\ 0 \le M \le 1, \forall u \in \mathcal{S},\ \ p_X > M p_X^u p_X^{u^c}$$

où p_Z est la densité de probabilité du vecteur aléatoire $\boldsymbol{Z}$, $u^c = \{1,\dots,p\} \setminus u$.
En dimension 2, on peut vérifier que cette condition implique des formes particulières de dépendance. La formule de Hoeffding reste valable sous ces conditions et on définit la sensibilité $S^{\boldsymbol{X}^u}$ pour le groupe $\boldsymbol{X}^u$ par :

$$\mathbf{Var}(Y)S^{\boldsymbol{X}^u} = \mathbf{Var}(\eta_u) + \sum_{u\cap v\neq u,v} \mathbf{Cov}\left(\eta_u(\boldsymbol{X}^u), \eta_v(\boldsymbol{X}^v)\right) \tag{I.2}$$

et en particulier pour la sensibilité du premier ordre :

$$\mathbf{Var}(Y)S^{X^i} = \mathbf{Var}(\eta_i(X^i)) + \sum_{i\notin v} \mathbf{Cov}\left(\eta_i(\boldsymbol{X}^i), \eta_v(\boldsymbol{X}^v)\right) \tag{I.3}$$

$\forall i = 1,\dots,p$.
On a donc $\sum_{u\in\mathcal{S}} S^{\boldsymbol{X}^u} = 1$.

Un cas particulier où cette méthode s'applique de façon évidente est le cas de lois bidimensionnelles du type $p(x^1,x^2) = \alpha h_1(x^1)h_2(x^2) + (1-\alpha)g(x^1,x^2)$ les marginales étant h_1, h_2. Pour p assez grand, la mise en œuvre nécessite des procédures d'orthogonalisation complexe [17]. Une fois la méthode posée son application nécessite des approximations par des développements limités dans un système orthogonal comme dans le cas indépendant.

4.2 Méta-modèles associés à des copules

La notion de copule est très utilisée dans le domaine de la finance. Elle permet surtout d'avoir un formalisme intéressant pour les extrêmes multivariées de séries chronologiques et fait l'objet de milliers de publications dans divers domaines. Les livres fondamentaux sont ceux de Sklar

[106] et Nelsen [88].
En terme de sensibilité, l'introduction de la notion de copule est naturelle à cause de la façon de modéliser l'incertitude sur les paramètres physiques. Dans les situations que nous évoquerons dans la dernière partie, l'incertitude sur le paramètre θ est modélisée par une loi uniforme sur l'intervalle $(\theta_{inf}, \theta_{max})$ et pour l'ensemble des paramètres θ_i par une loi portée par un parallélépipède $\Theta = \Pi_{i=1}^{p}(\theta_{i,inf}, \theta_{i,max})$.

Les incertitudes n'étant pas indépendantes, il s'agit de les modéliser par une loi sur Θ dont les marginales d'ordre 1 sont uniformes. Ces lois forment un ensemble convexe, compact en principe et décrit par ses points extrémaux. Ceux-ci, cependant, ne sont pas simples, peu pratiques car le support des points extrémaux est le plus souvent de dimension inférieure à p. D'autres éléments sont plus intéressants comme celui des lois d'entropie maximale. Les contraintes sur les marginales peuvent être vues comme des contraintes (en nombre infini) de type linéaire.

En pratique, à ces contraintes sur les marginales s'ajoutent des contraintes liées à la dépendance. Souvent on fixe la matrice de corrélation $\Gamma = \mathbf{Corr}(X^i, X^j)_{i,j=1,\dots,p}$, (il s'agit d'une contrainte de moments de type linéaire). On étudiera donc un sous ensemble de copules de matrice de corrélation Γ fixée.

Une copule de dimension d est une fonction de répartition C sur $[0,1]^d$ dont les marginales sont uniformes sur $[0,1]$.

Le théorème de représentation le plus général est celui de Sklar :

Proposition 3. — *Si C est une copule et $F_1, \dots, F_p$ des répartitions univariées alors pour tout $(x^1, \dots, x^p) \in R^p$:*

$$F(x^1, \dots, x^p) = C(F_1(x^1), \dots, F_p(x^p)) \tag{I.4}$$

et F est une fonction de répartition de $\mathbb{R}^p$ dont les marginales sont les fonctions $(F_1, \dots, F_p)$.

— *Réciproquement : Si F est une fonction de répartition de $\mathbb{R}^p$ dont les marginales sont les fonctions $(F_1, \dots, F_p)$ alors il existe une copule satisfaisant (I.4). Cette copule n'est pas unique mais si $(F_1, \dots, F_p)$ est continue alors C est unique et :*

$$\forall (u_1, \dots, u_p) \in [0,1]^p \;\; C(u_1, \dots, u_p) = F(F_1^{-1}(u_1), \dots, F_p^{-1}(u_p))$$

La copule la plus populaire est la copule Gaussienne associée à une distribution uniforme.
Soit $\boldsymbol{X} = (X^1, \dots, X^p)$ un vecteur Gaussien de corrélation Γ. D'après (I.4) on a :

$$U^i = \Phi(X^i) \tag{I.5}$$

où Φ est la répartition de la loi $\mathcal{N}(0,1)$ et $\boldsymbol{U} = (U^1, \dots, U^p)$ un vecteur aléatoire de composantes uniformes.
L'expression de la copule Gaussienne associée cette corrélation est :

$$C(x^1, \dots, x^p) = |\Gamma|^{-1/2} \exp(\frac{-1}{2}(\Phi^{-1}(u))(\Gamma^{-1} - I)(\Phi^{-1}(u))^*)$$

où le vecteur $\Phi^{-1}(u) = (\Phi^{-1}(u^i))_{i=1,\dots,p}$.

Le choix de la fonction de répartition ayant ses marginales uniformes est arbitraire mais la copule peut être contrainte par le choix de la corrélation Γ.

Soit $R = (R_{i,j})_{\substack{i=1,\dots,p \\ j=1,\dots,p}}$ la matrice de corrélation du vecteur $\boldsymbol{U}$, alors Γ la matrice de corrélation du vecteur $\boldsymbol{X}$ est obtenue à partir de R par :

$$\Gamma_{i,j} = 2\sin\left(\frac{\pi R_{i,j}}{6}\right) \tag{I.6}$$

La démonstration est donnée dans Biller et al. [8].

Une bonne approximation numérique peut être :

$$\Gamma_{i,j} = R_{i,j}(1,047 - 0.47(R_{i,j})^2) \tag{I.7}$$

La matrice R est donnée comme une matrice de corrélation mais rien ne garantit que la matrice Γ en soit une (matrice de type positif). On parle alors de problème réalisable [49].

Définition 3. *Soit $(F_i)_{i=1,\dots,p}$ une famille de fonctions marginales et R une matrice de corrélation. On dit que $((F_i)_{i=1,\dots,p}, R)$ est réalisable par une copule Gaussienne si et seulement si il existe un vecteur Gaussien $\boldsymbol{X}$ dont la matrice de covariance est Γ tel que :*

$$Z^i = (F_i)^{-1}\left(\Phi(X^i)\right), \quad i = 1, \dots, p$$

Dans le cas où la dimension de $\boldsymbol{X}$ est inférieure ou égale à 2, le problème est toujours réalisable. Ce qui n'est pas le cas en dimension supérieure. Si la matrice Γ n'est pas positive il est possible de trouver une matrice de corrélation proche de Γ [49].

L'avantage de ce type de méta-modèles pour calculer la sensibilité est que les fonctions de répartition étant des fonctions strictement croissantes les $\sigma-$ algèbres sont égaux :

$$\sigma(Z^i) = \sigma(X^i), \ i = 1, \dots, p \tag{I.8}$$

Alors :

$$S^{Z^i} = S^{X^i}$$

Nous détaillerons un exemple de calcul de sensibilité en dimension 2 au paragraphe 4.3.4 à partir de ce type de méta-modèles. Nous utiliserons d'autres copules uniformes dont les lois sont :

$$f(x, y) = \mathbb{1}_{\mathcal{C}}(x, y) + g(x)g(y)$$

avec $\mathcal{C}$ le carré $[0, 1]^2$, $\mathbb{1}_{\mathcal{C}}$ la fonction équivalente à $(x, y) \in [0, 1] \times [0, 1]$,
g fonction telle que $\int_0^1 g(x)\mathrm{d}x = 0$, $|g| \leq 1$ et $f(x, y) = 2\mathbb{1}_{0<x+y<1}\mathbb{1}_{\mathcal{C}}(x, y)$

4.3 Méthodes séquentielles : centrage et fonctions quantiles conditionnels

4.3.1 Centrage conditionnel

On suppose avoir une suite de variables $(X^1, \ldots, X^p)$ représentant le vecteur d'entrée $\boldsymbol{X}$. Dans ce paragraphe, on suppose la loi de $(X^1, \ldots, X^p)$ notée F et sa densité notée f connues. Les méthodes sont séquentielles. On se fixe un ordre sur les variables X^i. L'ordre est dicté par des considérations pratiques ou bien par des informations a priori sur l'importance des variables. Nous supposerons que c'est l'ordre naturel des variables sur les indices.

On note l'opérateur d'espérance conditionnelle par rapport à la σ-algèbre $\mathcal{B}_k = \sigma(X^1, \ldots, X^k)$:

$$\mathbf{E}(.|X^1, \ldots, X^k) = \mathbf{E}^{\lfloor 1,k \rfloor}(.)$$

On considère la suite $\bar{X}^k = X^k - \mathbf{E}^{\lfloor 1,k-1 \rfloor}(X^k)$.
On a :

$$\mathbf{E}(\bar{X}^k) = 0$$

.

Soit Z une variable $\mathcal{B}_{k-1}$ mesurable :

$$\mathbf{E}(Z\bar{X}^k) = \mathbf{E}(Z)\mathbf{E}(\bar{X}^k) = 0$$

Donc $\bar{X}^k$ est indépendante des fonctions $\mathcal{B}_{k-1}$ mesurables.

Dans [84], Mara et al. proposent d'étudier la sensibilité pour des entrées dépendantes à partir de cette transformation.
Si $Y = \eta(X^1, \ldots, X^p)$ alors on peut écrire :

$$Y = \eta(X^1, \bar{X}^2 + \mathbf{E}^{\lfloor 1,1 \rfloor}(X^2), \ldots, \bar{X}^p + \mathbf{E}^{\lfloor 1,p-1 \rfloor}(X^p))$$

Les nouvelles mesures de sensibilités sont données par :

$$\bar{S}^1 = \frac{\mathbf{Var}(\mathbf{E}(\eta(\boldsymbol{X})|\bar{X}^1))}{\mathbf{Var}(\eta(\boldsymbol{X}))}$$
$$\bar{S}^2 = \frac{\mathbf{Var}(\mathbf{E}(\eta(\boldsymbol{X})|\bar{X}^2))}{\mathbf{Var}(\eta(\boldsymbol{X}))}$$

On peut remarquer que $\bar{S}^1 = S^{X^1}$.

$\bar{S}^2$ est interprétée comme la contribution de X^2 diminuée de la contribution de X^1 et ainsi de suite. Nous ferons plus loin un retour sur cette méthode à la suite de celle que nous introduisons maintenant.

4.3.2 Transformation par la fonction quantile et application de la méthode Pick and Freeze aux variables dépendantes

Le lemme suivant de Rosenblatt [96], reprend une idée de Paul Lévy qui a montré l'existence de la transformation de la loi de $(X^1, \ldots, X^p)$ en loi produit en utilisant la fonction quantile de F.

Nous supposerons que la loi F de $(X^1, \ldots, X^p)$ admet une densité $f(\boldsymbol{x})$ par rapport à la mesure de Lebesgue $\mathbb{R}^p$.

Pour des raisons de simplicité d'écriture des fonctions réciproques, nous noterons :

$$support(f) = \overline{\{\boldsymbol{x},\ f(\boldsymbol{x}) > 0\}}$$

$\bar{.}$ désignant ici la fermeture. Nous supposons $f(\boldsymbol{x}) > 0$ pour $\boldsymbol{x}$ appartenant à l'intérieur de $support(f)$.

Les fonctions de répartition conditionnelles $F_{X^1},\ F_{X^2|X^1},\ \ldots,\ F_{X^p|X^1,\ldots,X^{p-1}}$ seront notées : $F_1,\ F_{2|1},\ \ldots,\ F_{p|1,\ldots,(p-1)}$.
La fonction de répartition conditionnelle est définie par :

$$F_{X^k|X^1=x^1,\ldots,X^{k-1}=x^{k-1}}(x^k) = \mathbf{P}(X^k < x^k | X^1 = x^1, \ldots, X^{k-1} = x^{k-1}) \tag{I.9}$$

$F_{X^k|X^1=x^1,\ldots,X^{k-1}=x^{k-1}}$ est donc une fonction de $\mathbb{R} \to [0,1]$ strictement croissante sur $support(f) = (a,b)$, nulle pour $x^k \leq a$ et égale à 1 pour $x^k \geq b$, $-\infty \leq a < b \leq +\infty$.

Lemme 2. *Avec les notations précédentes, il existe une transformation T telle que :*
$\boldsymbol{U} = T(\boldsymbol{X})$ *et* $\boldsymbol{U} = (U^1, \ldots, U^p)$ *est un vecteur aléatoire dont les $p-$coordonnées sont indépendantes et identiquement distribuées, uniformes (ou de densité fixée par avance).*

T n'est pas unique.

Une transformation T peut être définie ainsi :

$$\begin{aligned} U^1 &= F_1(X^1) \\ U^2 &= F_{2|1}(X^2) \\ &\ldots \end{aligned}$$

Plus généralement on écrira :

$$U^k = F_{k|1,\ldots,(k-1)}(X^k),\quad \forall 1 \leq k \leq p \tag{I.10}$$

U^k est donc une fonction $(X^1, \ldots, X^k)-$mesurable.

La démonstration du lemme 2 (voir [96]) résulte simplement d'une intégration multiple.

Étudions maintenant explicitement la transformation $T^{-1} : [0,1] \to \mathbb{R}^p$.

Soit $\overleftarrow{F}_.$ la fonction réciproque de $F_.$, la fonction de répartition définie telle que $\overleftarrow{F}_. : [0,1] \to \mathbb{R}$ on a :

$$X^1 = \overleftarrow{F_1}(U^1) = h_1(U^1)$$
$$X^2 = \overleftarrow{F_{2|1}}(U^2|U^1) = h_2(U^2, U^1)$$
$$\dots$$

d'où $X^k = \overleftarrow{F_{k|1,\dots,k-1}}(U^k|U^1, \dots, U^{k-1})$

X^k est donc sous la forme $h_k(U^1, \dots, U^k)$.
Un point à noter pour la suite est que cette transformation est utilisée depuis longtemps pour simuler une loi multidimensionnelle de répartition ou de densité donnée. On simule donc $U^1, \dots, U^p$ et l'on calcule : $X^1 = \overleftarrow{F_1}(U^1), X^2 = \overleftarrow{F_{2|1}}(U^2|U^1), \dots$
C'est la méthode la plus générale de simulation de lois multivariées. D'autres méthodes existent pour des lois particulières comme les lois de Student, Pearson,..., données par des formules simples.

Supposons maintenant que le modèle s'écrive : $Y = \eta(\boldsymbol{X})$ et que nous voulions calculer l'indice de sensibilité S^{X^1} par la méthode Pick and Freeze.

On veut fixer X^1 et représenter Y sous la forme :

$$Y = \tilde{\eta}(X^1, \boldsymbol{W})$$

où $\boldsymbol{W}$ est une variable indépendante de X^1.

D'après le lemme précédent on peut ré-écrire Y de la façon suivante :

$$Y = \eta(X^1 = U^1, X^2 = h_2(U^1, U^2), \dots, X^p = h_p(U^1, U^2, \dots, U^p)) = \tilde{\eta}(U^1 = X^1, U^2, \dots, U^p) \tag{I.11}$$

avec $\boldsymbol{W} = (U^2, \dots, U^p)$ et $\boldsymbol{W}$ indépendant de U^1.

Pour appliquer la méthode Pick and Freeze il faut construire deux échantillons Y et Y' avec la variable X^1 gelée, c'est-à-dire $Y = \tilde{\eta}(X^1, \boldsymbol{W})$ et $Y' = \tilde{\eta}(X^1, \boldsymbol{W'})$. En remarquant que $U^1 = F(X^1)$ est fixée lorsque X^1 l'est, pour appliquer l'estimateur Pick & Freeze, il suffit de simuler N fois un échantillon $(U^2, \dots, U^p)$ de loi uniforme sur $[0,1]^{p-1}$. En effet :

$$Y = \eta(X^1 = U^1, X^2 = h_2(U^1, U^2), \dots, X^p = h_p(U^1, U^2, \dots, U^p)) = \tilde{\eta}(X^1, \boldsymbol{W})$$
$$Y' = \eta(X^1 = U^1, X^2 = h_2(U^1, U'^2), \dots, X^p = h_p(U^1, U'^2, \dots, U'^p)) = \tilde{\eta}(X^1, \boldsymbol{W'})$$

où $\boldsymbol{W} = (U^2, \dots, U^p)$.

On peut évidement étendre la méthode au cas de plusieurs variables. Fixer (X^1, X^2) par exemple est équivalent à fixer (U^1, U^2).

Donc on aura dans ce cas :

$$Y' = \eta(X^1(U^1), X^2(U^1, U^2), \dots, X^p(U^1, U^2, U'^3, \dots, U'^p)) = \tilde{\eta}(U^1, U^2, U'^3, \dots, U'^p)$$

et on peut définir la sensibilité due à l'interaction par rapport à (X^1, X^2) par :

$$S^{X^1, X^2} - S^{X^1} - S^{X^2}$$

et l'estimer par la méthode Pick and Freeze.

La formule de Hoeffding canonique pour des variables dépendantes :

On défini la formule de Hoeffding canonique de la manière suivante :

Définition 4. *L'égalité* $\eta(X^1,\dots,X^p)=\tilde{\eta}(U^1,\dots,U^p)$ *définit une seule fonction* $\tilde{\eta}$ *invariante par permutation des coordonnées.* $\tilde{\eta}$ *ne dépend donc pas de l'ordre dans lequel on a construit* $(U^1,\dots,U^p)$ *par des formules du type :* $U^1=g_1(X^1),\dots,U^p=g_p(X^1,\dots,X^p)$ *et les formules réciproques qui permettent d'obtenir* $\eta\left(X^1=h_1(U^1),\dots,X^p=h_p(U^1,\dots,U^p)\right)$.

$\tilde{\eta}$ *est donc une forme canonique associée à* η *et il en est de même de la formule de Hoeffding qui lui est associée :*

$$\tilde{\eta}(U^1,\dots,U^p)=\sum_{k=1}^{p}\sum_{i_1\neq\dots\neq i_k}\eta_{i_1,\dots i_k}(U^{i_1},\dots,U^{i_k}) \tag{I.12}$$

qui permet de travailler dans les conditions de variables indépendantes.

Traduite en termes de $X^1,\dots,X^p$, cette formule I.12 ne donne évidemment pas la formule de Hoeffding dépendante telle qu'elle a été introduite dans Chastaing et al. [17] comme nous l'avons vu en 2.3.2 an début de ce chapitre. Chaque ordre permet de calculer des termes du premier ordre du type $\eta_1(X^j)$ de la formule de Hoeffding dépendante à condition de commencer par la bonne variable X^j.

Le cas le plus significatif qui montre la difficulté d'utiliser les formules d'Hoeffding dépendantes est celui de variables échangeables traité par Peccati [91]. Peccati construit une suite de trois variables d'entrée échangeables, conditionnellement indépendantes par rapport à une variable $\boldsymbol{W}$, et une fonction η elle même symétrique par rapport aux trois variables qui n'admet pas de décomposition de Hoeffding. Peccati donne une condition nécessaire d'existence d'une décomposition Hoeffding dépendante portant sur les opérateurs d'espérance conditionnelle.

On ne voit pas bien le niveau de pathologie qu'implique la condition nécessaire de Peccati et la violation des conditions suffisantes de Chastaing et al. mais il s'agit là d'un thème ouvert et difficile à comprendre quand existe une formule de Hoeffding dépendante. Le résultat de Peccati montre en tout cas, qu'il n'existe pas de combinatoire permettant de passer de la formule de Hoeffding canonique à la formule de Hoeffding dépendante sans des conditions restrictives. La condition d'existence d'une densité strictement positive sur l'intérieur de son support étant peut être suffisante.

Comment appliquer numériquement la méthode Pick and Freeze pour des variables dépendantes

Le fondement de la méthode repose sur l'égalité des $\sigma-$algèbres pour tous k :

$$\sigma(U^1,\dots,U^k)=\sigma(X^1,\dots,X^k) \tag{I.13}$$

alors :

$$\mathbf{E}(.|U^1,\dots,U^k)=\mathbf{E}(.|X^1,\dots,X^k) \tag{I.14}$$

Si l'on veut calculer toutes les sensibilités S^{X^j}, il faudra donc utiliser un ordre commençant par j, soit p ordres, si $j=1,\dots,p$.

Si l'on veut calculer l'ensemble des $S^{X^j}, S^{X^j X^l}$ il faudra donc utiliser $\frac{p(p-1)}{2}$ ordres. Ceci est certainement une faiblesse de la méthode du point de vue numérique. Chaque fois qu'un ordre est fixé, il faut calculer de manière récursive p fonctions de répartitions conditionnelles soit faire p intégrales multiples dans $\mathbb{R}^p$ et $\mathbb{R}^k$, $k \leq p$. Il y a donc dans $\mathbb{R}^p$ et $\mathbb{R}^k$ p^3 intégrales multiples à calculer si l'on veut toutes les sensibilités d'ordre 1 et 2.

Le grand intérêt de la méthode est que le calcul numérique inclut directement le processus de simulation de variables aléatoires dépendantes $(X^1, \ldots, X^p)$ à partir de variables aléatoires $i.i.d$ uniformes $(U^1, \ldots, U^p)$, elles, fournies par le simulateur. Sa récursivité est totale y compris pour le calcul des intégrales multiples récursives.

Il faut dans un premier temps choisir l'ordre dans lequel on souhaite calculer la sensibilité.

A partir de la densité on calcule les fonctions de répartition conditionnelles :
$F_1(X^1), F_{2|1}(X^2), \ldots, F_{p|1,\ldots,p-1}(X^p)$

Algorithm 1 Méthode transformation quantile

Require: N,$F_{i|1,\ldots,(i-1)}$ pour tout $i = 1, \ldots, p$

1: $U^1 = matrix(0, ncol = p, nrow = N)$; $U^2 = matrix(0, ncol = p, nrow = N)$
2: $X^1 = matrix(0, ncol = p, nrow = N)$; $X^2 = matrix(0, ncol = p, nrow = N)$
3:
4: $\boldsymbol{U^1} \sim \mathcal{U}$ et $\boldsymbol{U^2} \sim \mathcal{U}$
5:
6: $X^1[1,] =$ solution de $F_1(X) = U^1[1]$
7: $X^2[1,] = X^1[1,]$
8: **for** $i = 2$ to p **do**
9: **for** $j = 1$ to N **do**
10: $X^1[i,j] =$ solution de $F_{i|1,\ldots,(i-1)}(X) = U^1[i,j]$
11: $X^2[i,j] =$ solution de $F_{i|1,\ldots,(i-1)}(X) = U^2[i,j]$
12: **end for**
13: **end for**
14:
15: $Y_1 = \eta(X^1)$
16: $Y_2 = \eta(X^2)$
17: **return** Y_1, Y_2

L'algorithme complet de calcul de la sensibilité est le suivant :

— Simuler les $2N$ échantillons $(U^{1,(i)}, U^{2,(i)}, \ldots, U^{p,(i)})$ et $(U^{1,(i)}, \tilde{U}^{2,(i)}, \ldots, \tilde{U}^{p,(i)})$, où $(\tilde{U}^{2,(i)}, \ldots, \tilde{U}^{p,(i)})$ est une copie indépendante de $(U^{2,(i)}, \ldots, U^{p,(i)})$ (ligne 4).

— Résoudre récursivement en k les équations à une dimension d'inconnue $x^k = X(\boldsymbol{w})$ données par :

$$F_{X^{k,(i)}|X^{1,(i)}=x^1,\ldots,X^{k-1,(i)}=x^{k-1}} = U^{k,(i)}(\boldsymbol{w})$$

et de même pour $\tilde{U}^{k,(i)}(\boldsymbol{w})$.
La résolution aux lignes 6,10,11 se fait par exemple par la méthode de Newton ou encore par la méthode des directions alternées facile à appliquer ici puisque les fonctions $F_{i|1,\ldots,i-1}(x)$ sont monotones et continues si l'on ne sait pas calculer de manière explicite

les fonctions réciproques des fonctions de répartition conditionnelles. Numériquement la méthode n'est pas très coûteuse puisqu'elle est séquentielle et qu'une fois une grille choisie, il faut résoudre simplement l'équation :

$$F(x^n) = y^n$$

pour chaque point de la grille avec une très bonne valeur initiale.

— A l'aide de $(U^{1,(i)}, U^{2,(i)}, \ldots, U^{p,(i)})$ on fabrique $(X^{1,(i)}, X^{2,(i)}, \ldots, X^{p,(i)})$ et à l'aide de $(\tilde{U}^{2,(i)}, \ldots, \tilde{U}^{p,(i)})$ on fabrique $(\tilde{X}^{2,(i)}, \ldots, \tilde{X}^{p,(i)})$ (ligne 8 à 13).

— Par simulation ou calcul on obtient :

$$Y^{(i)} = \eta(X^{1,(i)}, X^{2,(i)}, \ldots, X^{p,(i)}), \quad Y'^{(i)} = \eta(X^{1,(i)}, \tilde{X}^{2,(i)}, \ldots, \tilde{X}^{p,(i)})$$
$$Y'^{(i)} = \tilde{\eta}(U^{1,(i)}, U^{2,(i)}, \ldots, U^{p,(i)}), \quad Y^{(i)} = \tilde{\eta}(U^{1,(i)}, \tilde{U}^{2,(i)}, \ldots, \tilde{U}^{p,(i)})$$

— A partir de $Y^{(i)}, Y'^{(i)}$, on calcule l'estimateur de $\hat{S}_N^{X^1}$

Si l'on veut calculer les sensibilités en $X^2, \ldots, X^p$ il faut recommencer toute l'opération en prenant des ordres du type $(X^2, X^{j_2}, \ldots, X^{j_p}), \ldots, (X^p, X^{j_2}, \ldots, X^{j_p})$

Si l'on veut calculer toutes les sensibilités d'ordre 2, il faut choisir $\frac{p(p-1)}{2}$ ordres différents de façon à représenter tous les couples (X^a, X^b) soit des ordres de types $(X^a, X^b, X^{j_3}, \ldots, X^{j_p})$

4.3.3 Exemple d'application du lemme

Nous allons comparer le calcul des indices de sensibilité directement puis en utilisant la transformation en variables indépendantes proposée précédemment.

On considère le modèle $\eta(X^1, X^2) = X^1 + X^2$, avec (X^1, X^2) un couple de variables aléatoires de loi uniforme sur le triangle (figure : I.1) :

$$D = \begin{cases} 0 \leq x^1, x^2 \leq 1 \\ x^1 + x^2 \leq 1 \end{cases} \tag{I.15}$$

La densité de probabilité du couple est : $f(x^1, x^2) = 2\mathbb{1}_D$, $\mathbb{1}_D$ équivaut à (x^1, x^2) définit sur le triangle D.

Explicitons d'abord les transformations permettant d'obtenir le couple de variables indépendantes (U^1, U^2). Nous avons besoin pour cela des fonctions de répartition F_1 et $F_{2|1}$.

La fonction de répartition F_1 de la variable X^1 est :

$$F_1(x^1) = (2x^1 - (x^1)^2), \quad x^1 \in [0, 1]$$

La fonction de répartition $F_{2|1}$ est :

$$F_{2|1}(x^2) = \frac{x^2}{1 - x^1}, \quad x^2 \in [0, 1 - x^1]$$

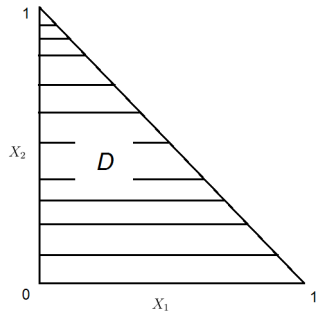

FIGURE I.1 – Domaine de définition de la densité de probabilité de la loi uniforme triangulaire

On en déduit les variables (U^1, U^2) :

$$U^1 = F_1(X^1) = 2X^1 - (X^1)^2$$
$$U^2 = F_{X^2|X^1}(X^2) = \frac{X^2}{1-X^1}\mathbb{1}_{[0,1-X^1]}$$

et les transformations inverses qui permettent de revenir aux variables initiales :

$$X^1 = \overleftarrow{F}(U^1) = 1 - \sqrt{1-U^1}$$
$$X^2 = U^2\sqrt{1-U^1}$$

Ce type de calcul explicite n'est pas possible en général.

Calculons la sensibilité par rapport à X^1 lorsque $\eta(X^1, X^2) = X^1 + X^2$ de deux manières différentes afin de voir comment fonctionne le changement de variables. Nous allons calculer :

$$S^{X^1} = \frac{\mathbf{E}(\eta(X^1, X^2)|X^1)}{\mathbf{Var}(\eta(X^1, X^2))} \quad \text{puis} \quad S^{X^1} = \frac{\mathbf{E}(\tilde{\eta}(U^1, U^2)|U^1)}{\mathbf{Var}(\tilde{\eta}(U^1, U^2))}$$

La densité de $(X^1 + X^2)$ est donc $2v\mathbb{1}_{0\leq v\leq 1}$ de variance $1/18$ obtenue à partir du changement de variable $(u = x, v = x + y)$.

Calcul pour les variables (X^1, X^2)

On a $\mathbf{E}(\eta(X^1, X^2)|X^1) = X^1 + \mathbf{E}(X^2|X^1) = X^1 + \displaystyle\int_0^{1-X^1} \frac{x^2}{1-X^1}\mathrm{d}x^2 = \frac{1+X^1}{2}$

La densité de X^1 étant $2(1-x^1)\mathbb{1}_{0\leq x^1+x^2\leq 1}$ on a donc :

$$\mathbf{Var}(\mathbf{E}(\eta(X^1, X^2)|X^1)) = \mathbf{Var}(\frac{1+X^1}{2}) = 1/72$$

L'indice de sensibilité du premier ordre est donc : $S^{X^1} = S^{X^2} = 1/4$

Calcul pour les variables (U^1, U^2)

L'expression de $\tilde{\eta}$ est par ailleurs :

$$\tilde{\eta}(U^1, U^2) = \eta(X^1, X^2) = 1 - \sqrt{1 - U^1} + \sqrt{1 - U^1} U^2$$

La sensibilité par rapport à la variable X^1 équivaut à calculer l'espérance conditionnelle de $\tilde{\eta}$ par rapport à U^1 :

$$\begin{aligned} \mathbf{E}(\tilde{\eta}|U^1) &= 1 - \sqrt{1 - U^1} + \frac{1}{2}\sqrt{1 - U^1} \\ &= 1 - \frac{1}{2}\sqrt{W} \end{aligned}$$

où W est uniforme.

Donc $\mathbf{Var}(\tilde{\eta}|U^1) = 1/72$

et donc $S^{U^1} = S^{X^1} = 1/4$

La sensibilité de S^{U^2}, n'a pas d'interprétation simple en termes de X^1, X^2 puisque $X^2 = U^2\sqrt{1 - U^1}$. Si nous calculons S^{U^2} nous trouvons :

$$\begin{aligned} \mathbf{E}(\tilde{\eta}|U^2) &= 1 + (U^2 - 1)\frac{2}{3} \\ \mathbf{Var}(\tilde{\eta}|U^2) &= 1/27 \end{aligned}$$

et donc $S^{U^2} = 2/3$ qui est différent de S^{X^2}. Pour avoir la sensibilité S^{X^2}, il faut réordonner le couple (x^1, x^2) en (x^2, x^1) pour trouver une transformation permettant d'obtenir un couple (U_2', U_1') de variables indépendantes.

Si l'on souhaite écrire la décomposition de Hoeffding en $U^1,\ U^2$ il faut et il suffit de centrer les variables $\sqrt{1 - U^1}$ et U^2 ce qui équivaut à :

$$\tilde{\eta}(U^1, U^2) = \frac{1}{3} - \frac{1}{2}(\sqrt{1 - U^1} - \frac{2}{3}) + \frac{2}{3}(U^2 + \frac{1}{2}) + (U^2 - \frac{1}{2})(\sqrt{1 - U^1} - \frac{2}{3})$$

Pour interpréter la formule de Hoeffding directement en termes de (X^1, X^2) nous renvoyons au travail de Chastaing et al. [17].

4.3.4 Application aux copules d'ordre 3 et modèle d'Ishigami

Considérons le modèle dit d'Ishigami :

$$Y = \sin(X^1) + 7\sin(X^2) + 0.1(X^3)^4 \sin(X^1) \tag{I.16}$$

où X^1 et X^3 ont pour corrélation ρ.
X^2 est indépendante de X^1 et X^3.
Les variables sont uniformes sur $[-\pi, \pi]$ alors que dans le calcul précédent elles étaient entre $[0, 1]$.

Nous considérons deux copules :

— une copule Gaussienne
— la copule donnée par la densité de probabilité :

$$f_\alpha(x,y) = \frac{1}{4\pi^2}\mathbb{1}_{[-\pi,\pi]^2}(x,y) + \alpha xy$$

Cas 1 : Copule Gaussienne

$Z^i, i = 1,2,3$ des variables Gaussiennes, on a alors :

$$X^i = (2\phi(Z^i) - 1) * \pi \text{ pour } i = 1,2,3$$

où Z^2 est indépendante de Z^1 et Z^3.

La corrélation ρ' de Z^1 et Z^3 est donnée par la relation :

$$\rho = \rho'(1.047 - 0.47(\rho')^2)$$

d'après les résultats de la section 4.2.

Si l'on cherche la sensibilité par rapport à X^1, les variables étant dépendantes et Gaussiennes, on utilise pour séparer les variables en variables indépendantes la méthode présentée dans la partie dynamique 2.2 appliquée à la variable Z^3 :

$$Z^3 = \rho' Z^1 + \sqrt{1-(\rho')^2}W \text{ où } W \sim \mathcal{N}(0,1)$$

Pour calculer la sensibilité, on utilisera la réécriture du système entrée sortie à partir des variables indépendantes de loi normale (Z^1, Z^2, W) :

$$Y = \sin(\pi(2\Phi(Z^1)-1)) + 7\sin(\pi(2\Phi(Z^2)-1)) + 0.1\left(\pi(2\Phi(\rho' Z^1 + \sqrt{1-\rho'^2}W) - 1)\right)^4 \sin(\pi(2\Phi(Z^1)-1))$$

On applique ensuite la méthode Pick and Freeze pour calculer la sensibilité. Les résultats sont donnés pour différentes valeurs de ρ sur la figure : I.2.

Suivant les valeurs de ρ les indices de sensibilité ne sont pas les mêmes. Le cas $\rho = 0$ correspond au cas où les variables sont indépendantes et l'on retrouve bien les résultats attendus (table : 4.1). Dans le cas $\rho = 1$, $X^1 = X^3$, les indices de sensibilité sont bien égaux.

Cas 2 : Copule f_α

On peut utiliser pour (X^1, X^3) des copules d'une autre forme par exemple de type :

$$f_\alpha(x,y) = \frac{1}{4\pi^2}\mathbb{1}_{[-\pi,\pi]^2}(x,y) + \alpha xy$$

Cette dernière n'existe pas pour toutes les valeurs de ρ fixées à l'avance. En effet pour s'assurer que f soit positive on doit avoir $|\alpha| \geq \frac{1}{4\pi^2}$, or $\mathbf{E}(X^1X^2) = \frac{\alpha 4\pi^3}{9}$, cela impose que $|\rho| \geq \frac{\pi}{9}$.

	ρ	S^{X^1}	S^{X^2}	S^{X^3}
copule f_α	10^{-7}	0.31	0.44	0
copule f_α	$\pi/9$	0.69	0.21	0.53
copule gaussienne	0	0.31	0.44	0
copule gaussienne	$\pi/9$	0.3	0.5	0.08

TABLE 4.1 – Calcul des sensibilités pour la copule gaussienne et f_α pour différentes valeurs de ρ

Supposons vouloir calculer la formule canonique pour la loi :

$$f_\alpha(x^1, x^3)\mathrm{d}x^1\mathrm{d}x^2\mathrm{d}x^3$$

On prend l'ordre $(1, 3, 2)$. On a toujours :

$$U^1 = \frac{X^1/\pi + 1}{2}$$
$$U^2 = \frac{X^2/\pi + 1}{2}$$

Reste à calculer U^3 :

$$U^3 = F_{X^3|X^1}(X^3) = \frac{1}{2\pi}(X^3 + \pi) + \pi\alpha X^1 \frac{(X^3)^2 - \pi^2}{2}$$

d'où :

$$X^3 = \frac{-1 + \sqrt{1 - 8\pi^2\alpha X^1(1 - \pi^2\alpha X^1 - 2U^3)}}{4\alpha\pi X^1}$$

et la formule canonique du modèle d'Ishigami pour cette loi sur (X^1, X^2, X^3) variables uniformes s'obtient en remplaçant :

$$X^1 \text{ par } (2U^1 - 1)\pi \tag{I.17}$$
$$X^2 \text{ par } (2U^2 - 1)\pi \tag{I.18}$$
$$X^3 \text{ par } \frac{-1 + \sqrt{1 - 8\pi^2\alpha X^1(1 - \pi^2\alpha X^1 - 2U^3)}}{4\alpha\pi X^1} \tag{I.19}$$

dans la sortie de Y. Pour calculer la sensibilité par rapport à X^3, on prend l'ordre 3,1,2 et c'est à X^1 que l'on applique la formule I.19.

Si l'on applique de manière numérique cette copule au modèle d'Ishigami et l'algorithme 1, on obtient les indices pour différentes valeurs de ρ figure : I.2.

Si l'on compare les deux copules pour différentes valeurs de ρ on ne trouve pas les même valeurs de sensibilité (table 4.1), à part pour $\rho = 0$ qui correspond au cas des variables indépendantes.

On peut remarquer que le choix d'une copule gaussienne par exemple est très arbitraire. S'il s'agit de paramètres physiques il est sans doute inacceptable parce que, bien que les marginales soient uniformes, le gradient près de la surface du cube est très fort.

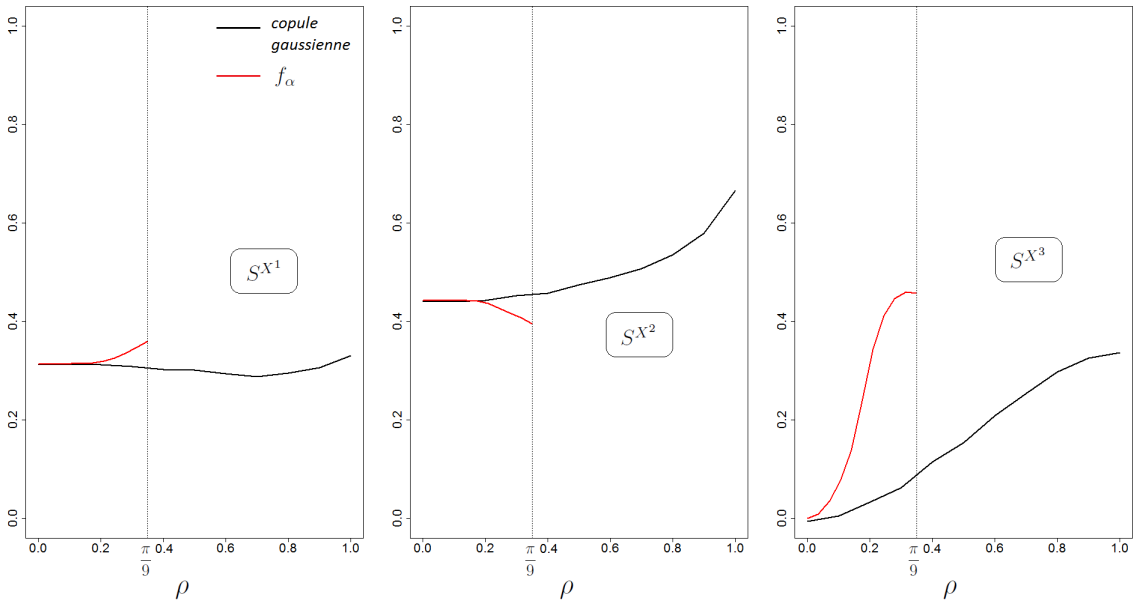

FIGURE I.2 – Indices de sensibilité pour différentes valeurs de ρ appliquées au modèle d'Ishigami pour la copule f_α et la copule gaussienne.

De plus le praticien doit avoir conscience que ces modèles sont "incomplets" lorsque l'on souhaite calculer la sensibilité. La corrélation ne représente pas de la bonne manière la dépendance entre les variables lorsque l'on souhaite calculer la sensibilité.

4.4 Méthode d'estimation non paramétrique

4.4.1 Estimation non paramétrique de la variance conditionnelle

Dans cette partie la loi F de l'entrée $\boldsymbol{X}$ n'est pas connue (ou mal connue) de même que la loi de sortie $Y = \eta(\boldsymbol{X})$. On cherche une estimation non paramétrique de $\mathbf{Var}(\mathbf{E}(Y|X^1))$ ou de quantiles analogues utiles pour définir la sensibilité. Les composantes de $\boldsymbol{X}$ ne sont pas supposées indépendantes. Du point de vue pratique comme usuellement, la qualité des méthodes dépendent de la faculté à simuler $\boldsymbol{X}$ et Y simultanément. Da Veiga et al. [25] supposent disposer d'un premier échantillon $(Y^{(i)}, \boldsymbol{X}^{(i)})_{i=1,\dots,N}$ et d'un deuxième échantillon de l'entrée $(\tilde{\boldsymbol{X}}^{(i)})_{i=1,\dots,N}$ indépendant de $(\boldsymbol{X}^{(i)})_{i=1,\dots,N}$. Le point de départ est l'utilisation d'un méta-modèle statistique, de type régression hétéroscédastique (de variances différentes) :

$$Y = m(X^1) + \sigma(X^1)\varepsilon$$

où ε est un bruit blanc.
On a $m(X^1) = \mathbf{E}(Y|X^1)$ donc (voir plus haut à propos des martingales) $Y - \mathbf{E}(Y|X^1)$ est indépendante de X^1. C'est aussi le cas de $\sigma(X^1)\varepsilon$ par hypothèse sur le méta-modèle, ε est indépendante de X_1 et centrée. Comment interpréter ε ? (les auteurs de [25] ne considèrent

pas le problème).
$X^{-1}-\mathbf{E}^{X^1}(X^{-1})$ est un vecteur indépendant de X^1. Si $X^{-1}=(X^2,\dots,X^p)$ on peut considérer ε comme de la forme $\psi(X^{-1}-\mathbf{E}^{X^1}(X^{-1}))$ qui est une fonction de la partie de X^{-1} indépendante de X^1. Dans [25], $m(\boldsymbol{x})$ et $\sigma^2(\boldsymbol{x})$ sont estimés par la méthode Loess qui y est détaillée. On peut ainsi estimer ces fonctions par splines ou par ondelettes. Avec le deuxième échantillon on a estimé empiriquement $\mathbf{Var}(\mathbf{E}(Y|X^1))$ et aussi $\mathbf{E}(\mathbf{Var}(Y|X^1))$. On pose :

$$\hat{T}_1=\frac{1}{N-1}\sum_{j=1}^{N}\hat{m}(\tilde{X}^{1,(j)}-\tilde{m})^2$$

$$\tilde{m}=\frac{1}{N}\sum_{j=1}^{N}m(\tilde{X}^{1,(j)})$$

$$\hat{T}_2=\frac{1}{N}\sum_{j=1}^{N}\hat{\sigma}^2(X^{1,(j)})$$

alors $\lim\limits_{N\to\infty}\hat{T}_1\to\mathbf{Var}(\mathbf{E}(Y|X^1))$ et $\hat{\sigma}_Y^2=\dfrac{1}{N}\sum\limits_{j=1}^{N}(Y^{(j)}-\bar{Y})^2$, d'où l'estimateur de S^{X^1} :

$$\hat{S}^1=\frac{\hat{T}_1}{\hat{\sigma}^2} \tag{I.20}$$

$\hat{\sigma}_Y^2-\hat{T}_2$ donne un second estimateur de $\mathbf{Var}(\mathbf{E}(Y|X^1))$.

4.4.2 Méta-modèle pour des entrées dépendantes. Estimation et choix des répartitions conditionnelles

Indiquons maintenant comment appliquer la méthode Pick and Freeze quand les lois sont inconnues, en estimant ces lois. Le nombre d'articles actuels concernant l'estimation des densités conditionnelles est considérable. Ce sujet est devenu central en économétrie pour choisir un modèle et faire des validations de modèle [25].

Nous venons de proposer une méthode pour appliquer l'estimateur Pick and Freeze au cas d'une entrée formée de variables aléatoires dépendantes. Il nous fallait partir de répartitions conditionnelles, calculables par des intégrations à partir de la densité de probabilité du vecteur d'entrée. Qu'en est-il si cette densité n'est pas connue, en particulier si l'aspect physique se restreint à des constantes par exemple le support de la loi du vecteur d'entrée ? On peut penser à estimer cette loi. Il y a un point de vue, disons classique, qui consiste à estimer la densité par exemple par une méthode non paramétrique (noyaux, LOESS, Fourier, LASSO ,... [60],[119],[73]). Cette méthode n'est pas bien adaptée ici car ces densités estimées dépendent beaucoup des paramètres de lissage choisis et sont souvent de qualité médiocre en dimension moyenne voir mauvaise en grande dimension. Nous proposons donc d'estimer directement des répartitions conditionnelles en dimension 1 de manière récursive. On trouvera le détail de la méthode utilisée dans l'article de Peter Hall et al.[53]. Nous choisissons une des méthodes exposées dans cet article, appelée méthode logistique.

Soit Y une variable aléatoire réelle et $\boldsymbol{X}$ un vecteur aléatoire à valeurs dans $\mathbb{R}^p$ (Y sera ici X^{k+1} et $\boldsymbol{X} = (X^1, \ldots, X^k)$. On dispose d'un échantillon $(Y^{(n)}, \boldsymbol{X}^{(n)})_{n=1,\ldots,N}$ et on veut estimer :

$$\Pi(y|\boldsymbol{x}) = P(Y < y|\boldsymbol{X} = \boldsymbol{x}) \tag{I.21}$$

Pour des raisons qui tiennent à la suite de ce travail, n pourra être aussi le temps et $(Y_n, \boldsymbol{X}_n)$ un processus stochastique stationnaire à temps entier.

On supposera $\boldsymbol{x} \to \Pi(y|\boldsymbol{x})$ r fois dérivable :

$$p(\boldsymbol{x}, \theta) = \theta_0 + \theta_1 \boldsymbol{x} + \cdots + \theta_{r-1}\boldsymbol{x}^{r-1}$$

et : $L(\boldsymbol{x}, \theta) = \dfrac{\exp(p(\boldsymbol{x}, \theta))}{1 + \exp(p(\boldsymbol{x}, \theta))}$

On choisira $\Pi(y|\boldsymbol{x})$ dans la famille définie par $L(\boldsymbol{x}, \theta)$.

La méthode dite logistique locale ou LOESS-logistique va consister à minimiser :

$$R(\theta, \boldsymbol{x}, y) = \left(\sum_{n=1}^{N} \mathbb{1}_{Y^{(n)} < y} - L(\boldsymbol{X}^{(n)} - \boldsymbol{x}, \theta) \right)^2 K_h(\boldsymbol{X}^{(n)} - \boldsymbol{x})$$

avec $K_h(.) = \dfrac{1}{h} K(\dfrac{\cdot}{h})$ où K est un noyau positif , $\displaystyle\int_{\mathbb{R}^k} K(\boldsymbol{x}) \mathrm{d}x = 1$.

On obtient alors un estimateur de $\theta = (\theta_0, \ldots, \theta_{r-1})$,

$$\hat{\theta}_N(\boldsymbol{x}, y) = \underset{\theta \in \mathbb{R}^k}{argmin} R(\theta, \boldsymbol{x}, y)$$

et par plug-in $\hat{\Pi}_n(Y|\boldsymbol{X}) = L(., \hat{\theta}(\boldsymbol{x}, y))$

$\hat{\Pi}_n(Y|\boldsymbol{X})$ est un estimateur convergent, le biais est de l'ordre de h^r et la variance de $\dfrac{1}{rh}$ avec $h = h(N) \to 0, \quad Nh(N) \to \infty$.

Du point de vue théorique, cette méthode vaut pour des séries chronologiques faiblement dépendantes telles que celles que nous utilisons dans le cas dynamique.

Dans P. Hall et al. [53], les auteurs suggèrent des procédures simples pour choisir r, l'ordre du modèle logistique, comme de partir d'un modèle polynomial de régression
$Y^{(i)} = a_0 + a_1 \boldsymbol{X}^{(i)} + \cdots + a_r (\boldsymbol{X}^{(i)})^r + \sigma \varepsilon^{(i)}$, avec résidus gaussiens et de choisir r en appliquant un critère de type Akaike. Pour choisir h_N, ils suggèrent d'utiliser un bootstrap, en définissant un jeu $(Y'^{(i)})_{i=1,\ldots,N}$ de nouvelles observations. A chaque jeu $(Y'^{(i)})_{i=1,\ldots,N}$ on associe une estimation :

$$\hat{\Pi}_n(Y'|\boldsymbol{X}) = \hat{\Pi}(Y'|\boldsymbol{X})$$

En posant $M(h, \boldsymbol{x}, y) = \mathbf{E}\left(\Pi_h(Y'|\boldsymbol{X}) - \tilde{\Pi}(Y'|\boldsymbol{X})|(\boldsymbol{X}_i, Y_i)\right)$, on choisit $h(\boldsymbol{X}, y)$ qui minimise M.

$\tilde{\Pi}(Y|\boldsymbol{X})$ est ici l'estimateur dit de Nadaraya-Watson de type local :

$$\tilde{\Pi}(Y|\boldsymbol{X}) = \frac{\sum_{i=1}^{N} \mathbb{1}(Y^{(i)} < y) p_i(\boldsymbol{X}) K_h(\boldsymbol{X}^{(i)} - \boldsymbol{X})}{\sum_{i=1}^{n} p_i(\boldsymbol{X}) K_h(\boldsymbol{X}^{(i)} - \boldsymbol{X})} \tag{I.22}$$

où $p_i(\boldsymbol{X})$ est un poids dépendant de $\boldsymbol{X}$.

Cet estimateur équivaut à un estimateur de type LOESS (voir Fan et Gybels [41]) localement linéaire.

On trouvera dans Hall et al. [53] des exemples de simulation par exemple pour des modèles dynamiques du type :

$$Y_t = 3.76Y_{t-1} + 0.235Y_{t-1}^2 + 0.3\varepsilon_t$$

Nous proposons d'utiliser le modèle logistique comme méta-modèle de la manière suivante pour des variables d'entrée dépendantes $(X^1, \ldots, X^p)$. On choisit un ordre $X^1, \ldots, X^p$ si on veut estimer les sensibilité S^{X^1} et $S^{X^1X^2}$ par exemple :

- Posons $U^1 = F(X^1)$. On estime F comme fonction de répartition classique à partir de l'échantillon $(X^{1,(i)})_{i=1,\ldots,N}$
- On estime $F(X^2|X^1)$ sous forme d'un modèle logistique
- On estime pas à pas $F(X^k|X^1, \ldots, X^{k-1})$ par $\hat{F}_{k|1,\ldots,k-1}$ sous forme logistique.

Le méta-modèle est constitué de p estimateurs d'entrée de répartitions conditionnelles $\hat{F}_{k|1,\ldots,k-1}$ sous forme logistique. On opère ensuite pour calculer la sensibilité comme précédemment lorsque ces répartitions sont connues.

Comparaison de la méthode avec la méthode de centrage conditionnel La méthode de centrage conditionnel du point de vue calcul nécessite un peu plus de calculs d'intégrales multiples mais pas de calcul de fonctions réciproques. Par contre, le calcul de la loi des variables indépendantes peut s'avérer assez compliqué et reste un handicap pour la simulation à la base de la méthode Pick and Freeze puisque l'on utilise le même calcul 2 fois : pour simuler $X^1, \ldots, X^p$ et pour estimer la sensibilité alors que pour la méthode des quantiles on ne la répète pas.

Chapitre 5

Conclusion

Nous avons d'abord présenté les méthodes les plus employées pour définir et calculer ou estimer la sensibilité. La définition que nous avons privilégiée est celle donnée par Sobol à partir de la variance de l'espérance conditionnelle de la sortie par rapport à une variable ou un paramètre d'entrée. Nous avons seulement indiqué le cadre général des différentes méthodes puisque l'objectif essentiel de cette thèse est ensuite de proposer une méthode adaptée au contexte dynamique des problèmes thermiques d'un bâtiment. Nous privilégierons dans la suite la méthode Pick and Freeze.

Les variables dans notre application seront dépendantes voire fortement dépendantes. La structure temporelle estimée nous amènera à privilégier dans un premier temps les structures de corrélation temporelle puis les propriétés de lois marginales. Dans la première partie, nous avons donc développé dans un cadre statique simple la méthode Pick and Freeze pour des variables dépendantes. Nous avons privilégié la méthode des quantiles conditionnels déjà connue pour simuler une loi de probabilité sur $\mathbb{R}^p$ quand elle n'est pas donnée par des formules closes très simples, cas assez rare. Cette méthode reste une des seules à pouvoir simuler des lois très générales. Elle n'est de fait concurrencée que par la recherche de chaines de Markov ayant la loi comme probabilité invariante. Cette méthode introduite par Paul Lévy consiste à transformer de manière récursive et non linéaire par une transformation mesurable et inversible T la loi en celle de p variables uniformes et indépendantes. Nous avons montré comment la méthode pouvait être adaptée au principe de Pick and Freeze, le point essentiel étant la conservation récursive de certaines sigma-algèbres. La méthode nécessite de prendre les variables dans un ordre donné et chaque ordre ne permet de calculer que certains indices de sensibilité. Pour calculer tous les indices d'ordre 1 on doit ainsi utiliser p ordres différents. C'est un handicap compensé à notre avis par des avantages importants. Il n'est pas utile d'utiliser la décomposition de Hoeffding ou des approximations de la sortie dans un système orthonormé pour calculer la sensibilité. Le calcul de T suffit. Ce calcul nécessite un minimum de calculs d'intégrales multiples mais aussi de fonctions réciproques. Tous ces calculs sont récursifs et celui faisant appel à des fonctions réciproques est très simple car il s'agit uniquement de fonctions définies sur $\mathbb{R}$ continues et surtout strictement monotones ce qui permet d'éviter si on le souhaite tout appel à la formule de Newton.

La méthode est illustrée sur des exemples. Dans le modèle d'Ishigami, nous prenons des va-

riables dépendantes de corrélation donnée. Il existe une infinité de modèles ayant des marginales uniformes et des corrélations données. Nous calculons par Pick and Freeze les sensibilités et montrons comment elles dépendent du modèle choisi, le modèle de copule gaussienne très utilisé pouvant être un mauvais choix. La méthode est certainement efficace pour comprendre les limites de la traduction de la dépendance en termes de corrélation.

L'ensemble de ces résultats sera utilisé dans les parties suivantes.

Deuxième partie

Outils probabilistes et statistiques adaptés à un cadre dynamique

Chapitre 1

Méta-modèles statistiques pour des phénomènes dynamiques

Une série chronologique ou temporelle à temps discret est un processus stochastique $(\boldsymbol{x}_t)$, $1 \leq t \leq n$, où t représente le temps (en minutes, jours, années ...). Les valeurs sont en général dans $\mathbb{R}^p$.
L'objectif de ce chapitre est de présenter les différentes notions et techniques permettant de comprendre les mécanismes qui génèrent ce type de données et de réaliser un modèle mathématique pour représenter ces mécanismes. Ces modèles peuvent être utilisés dans différents buts. Le but principal est la prévision ou le contrôle de systèmes à partir de ces prévisions ; mais un autre peut être de trouver un moyen de générer ces données. C'est dans ce dernier but que l'on souhaite les utiliser. Ayant privilégié la méthode Pick and Freeze et souffrant d'un manque important de données pour notre application nous allons rappeler les méthodes utilisées pour construire des métamodèles dynamiques et leur propriétés.

Plusieurs types de modèles linéaires existent : Auto-régressif (VAR), moyenne mobile (VMA), Auto-régressif à moyenne mobile ($VARMA$), etc... Ces modèles sont plus ou moins équivalents. Selon les domaines on préférera exprimer notre modèle en VAR ou VMA, le V signifiant vectoriel.
En dernière partie on propose d'exprimer les séries temporelles sous forme de modèles d'état. Ces modèles présentent un certain nombre d'avantages et sont très utilisés dans les domaines appliqués.

Nous utiliserons donc systématiquement dans ce travail des séries temporelles à des fins de modélisation des entrées des modèles entrée-sortie. Les observations de différents processus d'entrées (typiquement des températures ou le chauffage) et de sorties des systèmes linéaires ou non, sont modélisées par des processus dits du second ordre, le plus souvent Gaussiens. Nous examinerons aussi les cas non gaussiens et les propriétés qu'ils offrent

Il existe une littérature abondante à ce sujet. Nous ferons essentiellement référence aux livres de Brokwell & Davis ([13]), de Hannan [54] et d'Azencott & Dacunha Cateslle [4] pour les questions générales.

Dans le second chapitre nous proposerons des méthodes d'analyse de sensibilité adaptés à un cadre dynamique et entrées dépendantes.

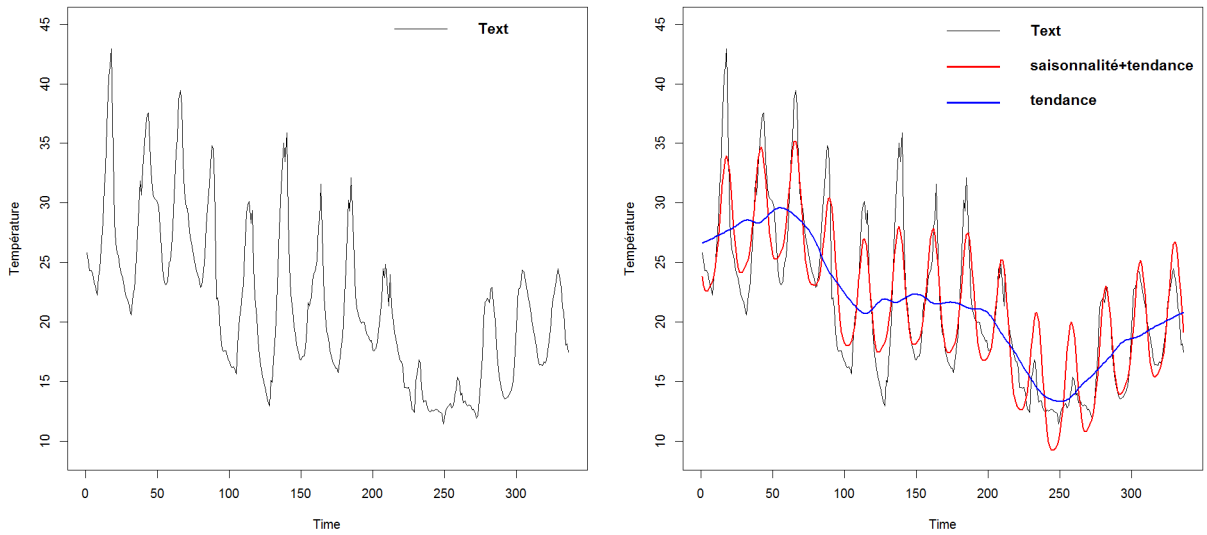

FIGURE II.1 – Exemple de série chronologique : Température extérieure en fonction du temps

1.1 Série centrée et série réduite : tendances et saisonnalités

La suite d'observation $(\boldsymbol{x}_t)$, $t \in \mathbb{N}$ ou $\mathbb{Z}$ peut être considérée comme une suite de réalisations de variables aléatoires indexées par $\mathbb{N}$ ou $\mathbb{Z}$. Ces séries sont modélisées par des processus $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$. Ces processus sont dits du second ordre, leur matrice de covariance étant finie.

L'espérance est notée :

$$m_t = \mathbf{E}(\boldsymbol{X}_t) \tag{II.1}$$

et la matrice de covariance :

$$\begin{aligned}\Gamma_X(t,t') &= (\mathbf{E}((X_t^i - m_t^i)(X_{t'}^j - m_t^j)))_{i,j=1,\dots,p} && \text{(II.2)}\\ &= \mathbf{E}((\boldsymbol{X}_t - m_t) \otimes (\boldsymbol{X}_{t'} - m_{t'})) && \text{(II.3)}\\ &= \mathbf{E}((\boldsymbol{X}_t - m_t)^*(\boldsymbol{X}_{t'} - m_{t'})) && \text{(II.4)}\end{aligned}$$

Nous n'emploierons qu'exceptionnellement la notation tensorielle $\otimes$.

On appelle série centrée, la série notée :

$$\boldsymbol{X}_t^c = \boldsymbol{X}_t - m_t$$

Dans le cas scalaire, on appelle série réduite la série suivante :

$$\boldsymbol{Z}_t = \frac{\boldsymbol{X}_t - m_t}{v_t} \quad \text{où} \quad v_t = \mathbf{E}((\boldsymbol{X}_t - m_t)^2)$$

On a plutôt l'habitude d'utiliser la série réduite en coordonnées :

$$\tilde{\boldsymbol{Z}}_t = \{Z_t^i = \frac{X_t^i - m_t^i}{v_t^i}, i = 1 \dots p\}$$

1.2 Décomposition en composantes saisonnières et tendance de la moyenne et de la covariance

Nous désignerons par tendance une composante déterministe d'une caractéristique du processus : moyenne, covariance, etc... qualitativement de basse fréquence.

Les phénomènes saisonniers sont inhérents aux problèmes appliqués qui nous concernent. Il s'agit de phénomènes présentant certaines périodicités, les périodes étant en général connues par avance.

Il existe deux approches pour modéliser ces phénomènes :

- l'approche purement stochastique utilisée en économétrie surtout et que nous n'utiliserons pas ici, approche dite SARIMA [54]
- l'approche déterministe utilisée en signal, climatologie et plus généralement en physique appliquée et que nous utiliserons.

Avant de chercher les liaisons explicatives pouvant exister entre les observations faites aux différents instants, il est nécessaire de corriger les données des saisonnalités et tendances. En effet, lors d'une période de croissance par exemple, les variables sont fortement corrélées entre elles sans que ceci exprime une quelconque liaison à caractère explicatif. On peut faire le même raisonnement pour la saisonnalité, il faut savoir si le phénomène est plus fort ou faible que d'habitude. C'est pourquoi il est souvent nécessaire de travailler sur les séries réduites et donc d'estimer les tendances et saisonnalités.

Si la série est saisonnière (ou s'il y a plusieurs saisonnalités : 24h, 7 jours, 365 jours, etc...), la moyenne m_t sera représentée sous la forme :

$$m(t) = S(t) + l(t)$$

où $S(t)$ est périodique de période T (éventuellement pluri-périodique) et $l(t)$ est une fonction lissée représentant un évolution basse fréquence.

La composante saisonnière S est déterminée par moindres carrés pénalisés à partir de l'observation sous forme de polynôme trigonométrique [12]. La saisonnalité sera en général étudiée composante X_t^i par composante et obtenue en minimisant :

$$\sum_{i=1}^{p} |X_t^i - S^i(t)|^2 - 2a(n)p^i$$

où p^i est le degré du polynôme trigonométrique $S^i(t)$. On prendra en général, $a(n) = log(n)$ ou $a(n) = constante$, le plus souvent 2. Il s'agit du critère d'Akaike introduit dans le chapitre 1.1.2.

Pour que la représentation de m soit identifiable, on fixera une contrainte, par exemple :

$$\int_0^T S(t)\mathrm{d}t = 0$$

La tendance $l(t)$ est alors estimée soit par un polynôme où les coefficients sont estimés par moindres carrés ou à partir de méthodes non paramétriques : méthodes à noyaux et plus proches voisins. Nous avons privilégié ici la méthode LOESS ([60]).

Les détails techniques de ces méthodes, utilisées ici pour des variables dépendantes dans un cadre non standard sont donnés dans [60].
Une fois $m(t)$ estimée, on opère de la même manière pour les variances : $\mathbf{E}(X_t^i - m^i(t))^2$ en les représentant sous la forme multiplicative $\mathbf{E}(X_t^i - m^i(t))^2 = V^i(t)s^i(t)$ où $V^i(t)$ est une tendance et $s^i(t)$ une fonction périodique.

Nous utiliserons occasionnellement les procédures pour la covariance $\Gamma_X(s,t)$ dont la dimension $(p \times p)$ rend les calculs très lourds. Néanmoins, le problème de saisonnalité est central pour $\Gamma_X(s,t)$ dans les applications que nous avons en vue. Nous développerons en fin de chapitre une approche particulière et adaptée : celle des processus vectoriels cyclo-stationnaires.

1.3 Propriétés fondamentales des processus stochastiques utilisés

La première propriété est la stationnarité.

Définition 5. *$(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ est dit stationnaire si et seulement si les lois multi-dimensionnelles de X sont invariantes par translation sur le temps, c'est-à-dire, pour tout (t,s) et tout $h > 0$, $\{\boldsymbol{X}_t, \boldsymbol{X}_{t+1}, \ldots, \boldsymbol{X}_{t+h}\}$ et $\{\boldsymbol{X}_s, \boldsymbol{X}_{s+1}, \ldots, \boldsymbol{X}_{s+h}\}$ ont la même loi.*

En particulier :

— la moyenne $\mathbf{E}(\boldsymbol{X}_t)$ est une constante m :

$$\mathbf{E}(\boldsymbol{X}_t) = m$$

— la covariance $\Gamma(t,s)$ ne dépend, pour tout (t,s), que de $(t-s)$:

$$\Gamma(t,s) = \gamma(t-s) = \Gamma_s$$

.

Si elle existe la densité spectrale ([13]) est donnée par :

$$f_X(\lambda) = \sum_{k\in\mathbb{Z}} e^{-ikt}\Gamma_k \tag{II.5}$$

La causalité par rapport au passé est une seconde propriété essentielle. Soit $V = \{V_\alpha, \alpha \in A\}$ une famille de variables aléatoires. Par $\mathrm{Span}(V)$ nous entendons l'espace de Hilbert engendré par les combinaisons linéaires finies d'éléments de V pour le produit scalaire :

$$< V_\alpha, V_\beta >= \mathbf{Cov}(V_\alpha, V_\beta) \tag{II.6}$$

On pose pour tout processus stochastique $(\boldsymbol{X}t)_{t\in\mathbb{Z}}$:

$$H_t^X = \mathrm{Span}(\boldsymbol{X}u, u \leq t) \tag{II.7}$$

$$H_{-\infty}^X = \bigcap_{t\in\mathbb{Z}} H_t^X \tag{II.8}$$

H_t^X est une suite croissante. Si $H_{-\infty}^X = \{0\}$, alors le processus $(\boldsymbol{X}t)_{t\in\mathbb{Z}}$ est dit causal. Il ne dépend pas d'un passé très lointain.

L'innovation à l'instant t est définie par :

$$\Omega_t^X = H_t^X \ominus H_{t-1}^X \tag{II.9}$$

Alors :

$$H_t^X = H_{t-1}^X \oplus \Omega_t^X \tag{II.10}$$

Si la dimension p de Ω_t^X est fixe alors le processus $(\boldsymbol{X}t)_{t\in\mathbb{Z}}$ est un processus $p-$vectoriel.

Lorsque $(\boldsymbol{X}t)_{t\in\mathbb{Z}}$ est stationnaire, nous dirons qu'il est de rang plein si $dim(\boldsymbol{X}t) = dim(\Omega_t^X) = p$ et donc qu'il existe une base de dimension p de Ω_t^X pour tout t.

En pratique nous nous placerons toujours dans ce cas qui est celui d'un processus stationnaire.

Soit Π_K l'opérateur de projection sur un sous espace K de H^X. $H^X = \lim_{t\to\infty} H_t^X$ alors $H_t^X = \underset{u=\infty}{\oplus} \Omega_u^X$.

Si X est stationnaire et causal :

$$\boldsymbol{X}_t = \sum_{u=0}^{\infty} A_u \boldsymbol{\omega}_{t-u} \tag{II.11}$$

où $(A_u)_{u\in\mathbb{N}}$ une suite de matrices de dimension $(p \times p)$ telles que : $\sum \|A_u\|^2 < \infty$, dite représentation de Wold en moyenne infinie.

1.4 Processus $\boldsymbol{VAR}(p)$ et $\boldsymbol{VARMA}(p, q)$

Le cas le plus simple de processus du second ordre linéaire est celui des processus $VAR(1)$. La représentation du processus VAR correspond à une description de l'évolution du système à partir des valeurs d'un nombre fini des observations passées.

Un processus $VAR(1)$ est défini par l'équation de récurrence :

$$\boldsymbol{X}_t = A\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t \tag{II.12}$$

où A est une matrice $(p \times p)$ que nous supposerons toujours non singulière et $\boldsymbol{\varepsilon}_t$ un bruit blanc vectoriel de dimension p.

Pour que l'équation admette une solution stationnaire il faut et il suffit que le rayon spectral de A (plus grand module des valeurs propres) soit inférieur à 1. La solution est alors unique. Nous supposons dans la plupart des applications que $(\boldsymbol{\varepsilon}_t)_{t\in\mathbb{Z}}$ est une suite de vecteurs aléatoires indépendants, équidistribués et le plus souvent Gaussiens.

On note Γ^ε la matrice de covariance de ε non dégénérée de rang p, qui vérifie la relation fondamentale :

$$(I - AA^*)\Gamma_t^X = \Gamma^\varepsilon \text{ avec } \Gamma^X = \Gamma_0^X \tag{II.13}$$

On a alors :

$$\Gamma_k = \mathbf{E}(\boldsymbol{X}_t^* \boldsymbol{X}_{t-k}) = A^k A^* \Gamma^X$$

Le modèle II.12 peut se réécrire sous la forme :

$$\boldsymbol{X}_t = \sum_{k=0}^{t-1} A^k \boldsymbol{\varepsilon}_{t-k} + A^* \boldsymbol{X}_0 \tag{II.14}$$

$$= \sum_{k=-\infty}^{t-1} A^k \boldsymbol{\varepsilon}_{t-k} \tag{II.15}$$

avec $\mathbf{E}\|A^* \boldsymbol{X}_0\|^2 \leq \|A\|\|\Gamma^X\|$.

Les processus $VAR(p)$ sont obtenus par des équations de récurrence du type :

$$\boldsymbol{X}_t = A_1 \boldsymbol{X}_{t-1} + \cdots + A_p \boldsymbol{X}_{t-p} + \boldsymbol{\varepsilon}_t \tag{II.16}$$

Soit d l'opérateur retard défini sur H^X par $d\boldsymbol{X}_t = \boldsymbol{X}_{t-1}$ (translation de -1 dans le temps). Soit $P(d)$ un polynôme de d. On peut écrire :

$$P(d)\boldsymbol{X}_t = \boldsymbol{\varepsilon}_t, \quad P(d) = I - \sum_{j=1}^{p} A_j d^j$$

La condition nécessaire et suffisante d'existence d'une solution stationnaire est :

$$\det(I_p - \sum_{j=1}^{p} A_j z^j) \neq 0 \text{ pour } z \leq 1$$

Par simplicité, on cherche à se ramener à un processus $VAR(1)$ en changeant les dimensions.

En partant du processus unidimensionnel X_t^j d'ordre p_j, on considère le vecteur :
$\boldsymbol{Z}_t^j = (X_t^j, \ldots, X_{t-p_j}^j)^*$ et l'équation :

$$X_t^j = \sum_{k=1}^{p_j} a_j^k X_{t-k}^j + \epsilon_t^j$$

s'écrit

$$Z_t^j = A_j Z_{t-1}^j + \xi_t^j$$

où

$$A_j = \begin{pmatrix} a_j^1 & a_j^2 & a_j^3 & \cdots & a_j^{p_j} \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \tag{II.17}$$

est la matrice dite compagnon du vecteur : $(a_j^1, \ldots, a_j^{p_j})^*$
$\xi_t^j = (\varepsilon_t^j, 0, \ldots, 0)$ le vecteurs des erreurs de dimension p_j.

Plus généralement si on considère un $VAR(p)$ du type : $X_t^j = \sum_{k=1}^{p_j} A_j^k X_{t-k}^j + \varepsilon_t^j$, on transforme le $VAR(p)$ en $VAR(1)$:

$$\boldsymbol{Z}_t = \tilde{A}\boldsymbol{Z}_{t-1} + \boldsymbol{\xi}_t$$

Avec :

$$\tilde{A} = \begin{pmatrix} A_1 & A_2 & A_3 & \cdots & A_p \\ I & 0 & 0 & \cdots & 0 \\ 0 & I & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & I & 0 \end{pmatrix} \tag{II.18}$$

L'existence d'une solution stationnaire est équivalente à $\|\tilde{A}\| \leq 1$.

Les processus $VARMA(p,q)$ sont des processus Auto-Régressif d'ordre p à Moyenne Mobile d'ordre q Vectoriels.

Les processus $VARMA(p,q)$ sont définis par une équation du type :

$$\Phi(d)\boldsymbol{X}_t = \Theta(d)\boldsymbol{\varepsilon}_t \tag{II.19}$$

où

$$\Phi(d) = I - \sum_{i=1}^{p} A_i d^i$$

$$\Theta(d) = I + \sum_{i=1}^{q} B_i d^i$$

et $\boldsymbol{\varepsilon}_t$ est un bruit blanc de covariance Γ^{ε}.

Il existe une solution stationnaire si et seulement si : $\|\Phi\| < 1$ et $H^{\varepsilon}_{-\infty,t} = H^{X}_{-\infty,t}$ si et seulement si $\|\Theta\| < 1$.

On a alors :

$$\boldsymbol{\varepsilon}_t = \Theta(d)^{-1}\Phi(d)\boldsymbol{X}_t$$

et la représentation $VAR(\infty)$ des processus $VARMA$

$$\sum_{i=0}^{\infty} \Psi_i \boldsymbol{X}_{t-i} = \boldsymbol{\varepsilon}_t$$

1.5 Introduction d'une co-variable : processus *VARX*

Les processus VAR ne permettent parfois que de faire une étude descriptive du problème et non explicative. De façon à introduire cet aspect explicatif, on considère d'autres variables pouvant avoir une influence sur les variables décrites par le processus considéré et dont les

valeurs sont fixées extérieurement au phénomène que l'on tente de décrire. Par exemple la température d'une pièce dépend de la température extérieure. La température extérieure est alors appelée variable exogène.

On définit alors les processus *VARX* (Vector AutoRegressive with eXogenous variable).

Un processus *VARX* est défini par une équation :

$$P(d)\boldsymbol{X}_t = C(d)\boldsymbol{Z}_t + \boldsymbol{\varepsilon}_t$$

où P et C sont deux polynômes et $\boldsymbol{Z}_t$ est un processus stationnaire causal.

De même un processus *VARMAX* sera défini par :

$$P(d)\boldsymbol{X}_t = Q(d)\boldsymbol{\varepsilon}_t + C(d)\boldsymbol{Z}_t \tag{II.20}$$

1.6 Processus cyclo-stationnaire

Il s'agit d'une extension de la notion de stationnarité à des phénomènes aléatoires dont la moyenne et la covariance sont périodiques, ce qui ne signifie pas que les trajectoires le sont.

Définition 6. *$(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ est dit cyclo-stationnaire si toutes les lois multidimensionnelles sont invariantes par translation τ sur le temps définie sur H^X par $\boldsymbol{X}_t \to \boldsymbol{X}_{t+h\tau}$, τ est la période du processus cyclo-stationnaire, $h \in \mathbb{Z}$.*

Pour un processus cyclo-stationnaire la moyenne $m(t)$ est une fonction périodique et la covariance Γ^X est telle que :

$$\Gamma^X(t + k\tau, t' + k\tau) = \Gamma^X(t, t')$$

Γ^X est donc associée aux $(\tau - 1)$ matrices définies par $s \mapsto \Gamma(t, t + s), \quad t = 0, \ldots, \tau - 1$

Pour un processus gaussien, cette périodicité de m et de Γ équivaut à la cyclo-stationnarité.

Le plus souvent on adopte une représentation vectorielle en posant :

$$\boldsymbol{Z}_t^X = (\boldsymbol{X}_{t\tau}, \ldots, \boldsymbol{X}_{t\tau+\tau-1})$$

Le processus $\boldsymbol{X}_t$ de dimension r est transformé en processus de dimension $r\tau$.

Nous utiliserons en particulier les processus $VAR(1)$ cyclo-stationnaires du type :

$$\boldsymbol{X}_t = A_{[t]}\boldsymbol{X}_{t-1} + \varepsilon_t \tag{II.21}$$

Avec $[t] = t \ \ mod(\tau)$ et $\|A_i\| < 1$ pour $i = 1, \ldots, \tau - 1$.

Théorème 3. *La cyclo-stationnarité de $\boldsymbol{X}_t$ est équivalente a la stationnarité du processus Z_t^X associé.*

1.7 Statistique des processus *VAR*, *VARMA*, *VARMAX*

Dans ce paragraphe, nous donnons les résultats que nous avons utilisés dans cette étude en particulier pour les processus $VAR(p)$.

1.7.1 Estimation paramétrique

Si $\boldsymbol{X}_t = A_1\boldsymbol{X}_{t-1} + \cdots + A_p\boldsymbol{X}_{t-p} + \varepsilon_t$ et si $\boldsymbol{Z}_t = (\boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_{t-p})^*$ alors l'estimateur des moindres carrés s'écrit :

$$\hat{A} = (\hat{A}_1, \ldots, \hat{A}_p) = (\sum_{t=1}^{n} \boldsymbol{Z}_t^* \boldsymbol{Z}_t)^{-1} (\sum_{t=1}^{n} \boldsymbol{X}_t \boldsymbol{Z}_t^*)$$

Il est systématiquement utilisé si on ne connait pas la loi des résidus. Il n'y a de bonnes propriétés asymptotiques que sous certaines conditions (voir [55],[13])

Si les résidus sont gaussiens, de covariance Γ_k, on utilise l'estimateur du maximum de vraisemblance. Pour un processus $VAR(1)$ la vraisemblance de $\boldsymbol{\varepsilon}_t$ est :

$$L_n(\boldsymbol{X}; p) = \frac{-n}{2} log(2\pi) - \frac{n}{2} log(\det(\Gamma^{\boldsymbol{\varepsilon}})) - \frac{1}{2} \sum_{t=1}^{n} \boldsymbol{\varepsilon}_t^* \Gamma^{\varepsilon} \boldsymbol{\varepsilon}_t$$

puisque les $\boldsymbol{\varepsilon}_t$ sont indépendants.

$\boldsymbol{\varepsilon}_t = (\boldsymbol{X}_t - A\boldsymbol{X}_{t-1})$ d'où la log-vraisemblance à une constante près :

$$-\frac{n}{2} log(\det(\Gamma^{\varepsilon})) - \frac{1}{2} \sum_{t=1}^{n} (\boldsymbol{X}_t - A\boldsymbol{X}_{t-1})^* \Gamma^{\varepsilon} (\boldsymbol{X}_t - A\boldsymbol{X}_{t-1}) \tag{II.22}$$

Sa maximisation permet d'estimer A et Γ^{ε}.

Pour un processus $VAR(p)$, on le transforme en $VAR(1)$ puis on calcule la vraisemblance comme précédemment. Dans le cas d'un processus $VARMA$ les choses sont plus compliquées et il existe plusieurs algorithmes d'estimation. Les algorithmes les plus simples et performants sont les algorithmes d'aller-retour dus à Box et Jenkins et dont la convergence est démontrée dans [13] pour le cas unidimensionnel. Dans le cas général on peut se référer à [55]. Les mêmes méthodes s'appliquent aux processus $VARMAX$.

1.7.2 Détermination de l'ordre p d'un processus VAR(p)

Supposons le processus défini par :

$$\boldsymbol{X}_t - A_1\boldsymbol{X}_{t-1} - \cdots - A_t\boldsymbol{X}_{t-p} = \boldsymbol{\varepsilon}_t$$

En général pour déterminer p, on utilise la vraisemblance $L(\cdot)$ pénalisée pour une suite de modèles VAR(p) emboîtés, lorsque p croit de 1 à p_0 par exemple.

La pénalisation pour (p, τ) varie en sens opposé de la log-vraisemblance quand p varie. Elle décroit (croit en valeur absolue) quand p croit et donc :

$$L(\boldsymbol{X}; p) + pen(p, n), \quad \text{où } n \text{ le nombre d'observations}$$

passe par un maximum.

Le critère le plus connu est le critère d'Akaike :

$$pen(p, n) = -2p$$

Ce critère n'est pas consistant quand $t \to \infty$, les estimateurs ne sont pas convergents.

Nous utiliserons le critère BIC en général qui est consistant et adapté aussi au point de vue bayésien :

$$pen(p, n) = -2p\ log(n)$$

En pratique la détermination globale de l'ordre peut être mal adaptée car elle procède par compensation entre les ordres des différentes composantes et ne vaut que pour n très grand. Il est donc préférable de contrôler l'ordre global par les ordres de chacune des composantes et de voir dans la matrice estimée $\hat{A}$:

$$(\hat{A}, \hat{p}, \hat{\Gamma^{\varepsilon}}) = \underset{A,p,\Gamma}{\textbf{argmax}}\ L_n(\boldsymbol{X}, p, A, \Gamma^{\epsilon}) - p\ log(n)$$

si certains termes peuvent être négligés.

Ces résultats s'étendent au cas cyclo-stationnaire. Si les dimensions sont importantes, une fois le processus cyclo-stationnaire déplié en $VAR(1)$, sa dimension devient $r\tau$, τ étant la période. Pour éviter l'explosion du nombre de paramètres, nous avons cherché à tester les coefficients que l'on peut annuler, mais ces pratiques restent empiriques et devront être développées.

Intervalle de confiance et loi de probabilité des paramètres :

Les estimateurs que nous venons de définir pour A et Γ^{ε} sont convergents et les erreurs d'estimation asymptotiquement normales. Comme nous travaillerons essentiellement sur des formes $VAR(1)$ stationnaires et cyclo-stationnaires, il est simple d'utiliser des méthodes de bootstrap pour estimer les lois des estimateurs.

Si $\hat{A}$ et $\hat{\Gamma^{\varepsilon}}$ sont des estimateurs obtenus par maximum de vraisemblance pénalisé notés :

$$\hat{\boldsymbol{\varepsilon}}_t = \boldsymbol{X}_t - \hat{A}\boldsymbol{X}_{t-1}$$

la non indépendance des $\boldsymbol{X}_t$ exige d'effectuer pour les $\boldsymbol{\varepsilon}_t$ un ré-échantillonnage par blocs ou tout autre méthode justifiant le bootstrap.

Le ré-échantillonnage se fait en tirant au sort avec remise les $\hat{\boldsymbol{\varepsilon}}_t$ pour obtenir les $\boldsymbol{\varepsilon}_t^{(i)}$, i correspondant au i^{eme} tirage au sort. Pour tenir compte de la dépendance, le tirage au sort des $\boldsymbol{\varepsilon}_t^{(i)}$ se fait par blocs, ici de longueur 2 ou 3 pour un $VAR(1)$ des valeurs successives des $\hat{\boldsymbol{\varepsilon}}_t$. Il existe plusieurs variantes de la construction des blocs qui peuvent être soit disjoints, soit glissants.

1.8 Statistique des processus cyclo-stationnaires $VARCS$

On considère un processus cylco-stationnaire de la forme :

$$\boldsymbol{X}_t = A_{[t]}\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_{t,[t]}$$

où $[t] = t \mod(\tau)$.
$(\boldsymbol{\varepsilon}_{t,[t]})$ est un ensemble de r variables aléatoires i.i.d pour $[t] = 0, 1, \ldots, \tau$ fixé.
$A_{[t]}$ est une matrice $r \times r$ dont les coefficients sont périodiques. Ils seront donc représentés par des fonctions $A_{i,j}(t)$, dont le nombre de coefficients peut dépendre de $[t]$ mais qui en général sera pris en respectant le principe de parcimonie, éventuellement par un critère type Akaike.

On a vu qu'un processus cyclo-stationnaire de dimension p se ramène à un processus stationnaire de dimension $p \times p$. Chaque fois que possible, on utilisera cette transformation avant d'estimer les paramètres des matrices $A_{[t]}$. La difficulté pratique est donc éventuellement de remplacer les constantes $A_{i,j}$ intervenant dans un processus stationnaire par les fonctions périodiques $A_{i,j}[t]$. L'ordre pratique du processus sera donc donné par nombre de coefficients de $A_{i,j}(t)$ si on choisit un degré fixe pour les fonctions $A_{i,j}(t)$. Ceci sera illustré dans la partie III.

1.9 Exemple de traitement d'une série chronologique appliquée à la température extérieure

La température extérieure, notée T^{ext}, peut ne pas être considérée comme une variable d'entrée à proprement parler mais un processus d'excitation des autres températures présentes dans le bâtiment. C'est pourquoi nous avons décidé de présenter en exemple la modélisation de T^{ext} de manière univariée.

La température extérieure est mesurée heure par heure du 22 Août au 9 Septembre 2011.

La tendance est estimée par une méthode non paramétrique de type LOESS.

La saisonnalité est estimée par un polynôme trigonométrique de période 24 heures. Les coefficients sont estimés par une méthode de moindres carrés et le degré sélectionné est déterminé en minimisant le critère de AKAIKE. La tendance et la saisonnalité obtenues sont tracées sur la figure II.2.

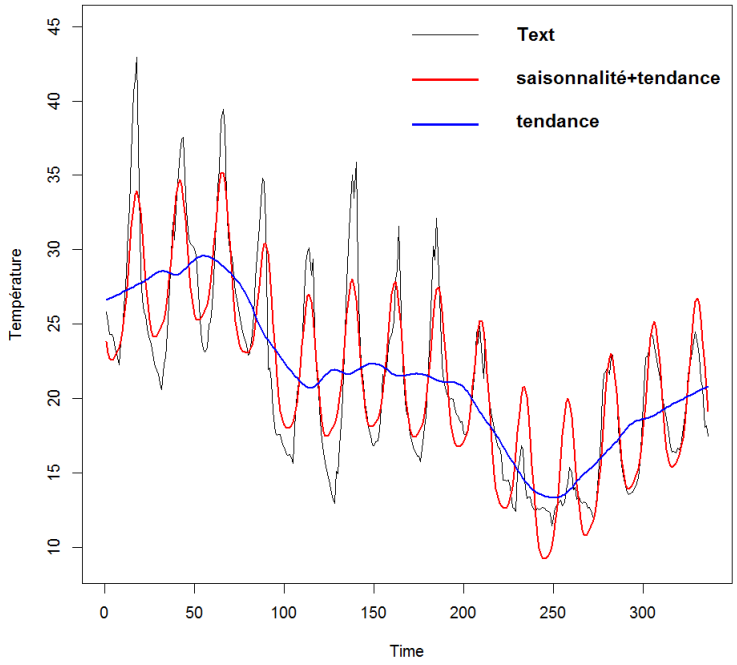

FIGURE II.2 – Température extérieure en fonction du temps mesurée

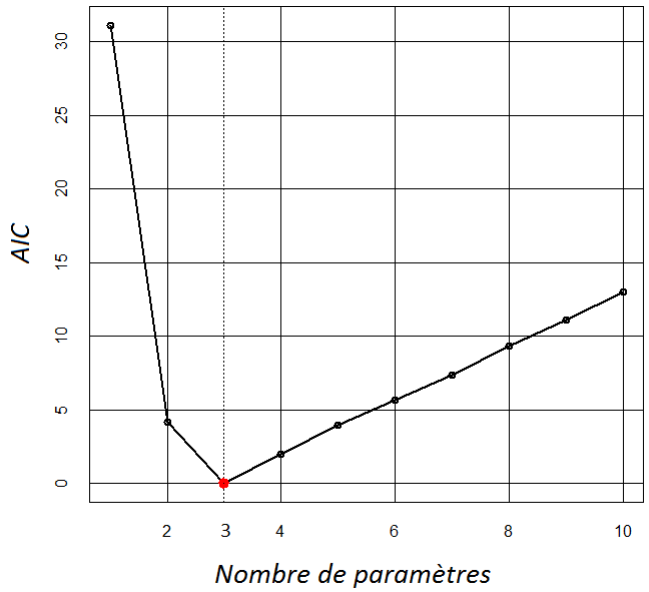

FIGURE II.3 – Critère AIC en fonction du nombre de variables retenues pour le modèle AR

Après traitement, c'est-à-dire après avoir retiré les tendances et les saisonnalités, nous ajustons un modèle AR aux données stationnarisées (figure II.4) estimé par maximum de vraisemblance. Le modèle retenu par critère AKAIKE est un modèle $AR(3)$ (figure : II.3).

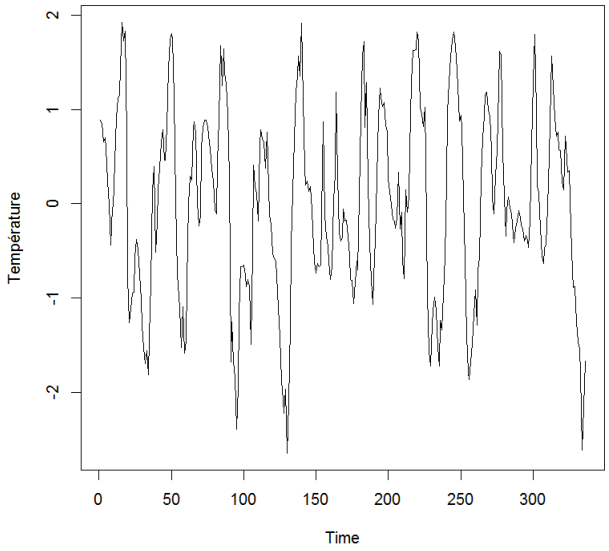

FIGURE II.4 – Température extérieure après retrait des saisonnalités et de la tendance en fonction du temps

$$T_t^{\text{ext}} = 1.11T_{t-1}^{\text{ext}} - 0.13T_{t-2}^{\text{ext}} - 0.13T_{t-3}^{\text{ext}} + \varepsilon_t$$

$\varepsilon_t \sim \mathcal{N}(0, 1.59)$

Sur la figure II.5, le modèle semble bien suivre les données mesurées. Pour contrôler le modèle retenu on peut tester les hypothèses de normalité sur les résidus et leur indépendance.

Pour étudier la structure de dépendance on trace la fonction d'autocovariance (ACF) (figure II.7) définie par :

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} \text{ avec } \gamma(h) = \mathbf{Cov}(\boldsymbol{X}_{t+h}, \boldsymbol{X}_t) \tag{II.23}$$

Si les observations sont indépendantes entre elles alors :

$$\gamma(t+h, t) = \begin{cases} \sigma^2, & si \;\; h = 0 \\ 0 & sinon \end{cases}$$

On peut remarquer aussi que si les données sont stationnaires alors :

$$\begin{aligned} &\gamma(0) \geq 0 \\ &\|\gamma(h)\| \leq \gamma(0) \quad \forall h \in \mathbb{Z} \\ &\gamma(h) = \gamma(-h) \quad \forall h \in \mathbb{Z} \end{aligned}$$

On constate bien que seule $\gamma(0)$ est important, alors les résidus sont indépendants et stationnaires.

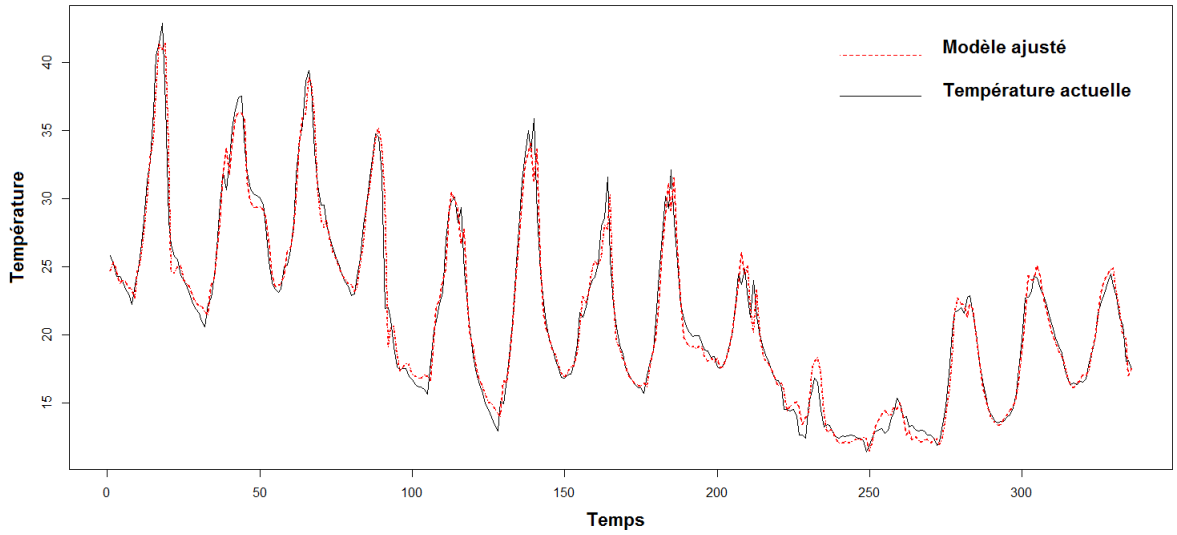

FIGURE II.5 – Modèle AR ajusté aux données

Quand à la normalité on trace le diagramme quantile-quantile (qq-plot) figure II.6. On compare la position de certains quantiles dans la population observée avec leur position dans la population théorique. Si la courbe est une droite cela signifie que les résidus suivent une loi normale.

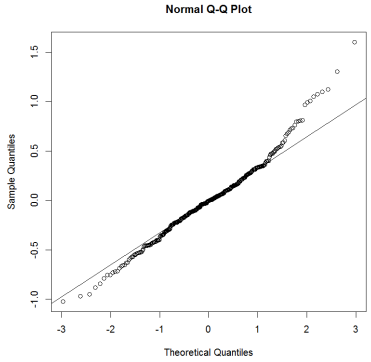

FIGURE II.6 – Tracé qq-plot des erreurs

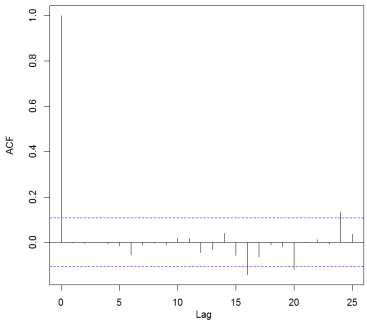

FIGURE II.7 – Tracé de la fonction d'Autocorrélation des erreurs

Ici, figure II.6, nous obtenons une droite. Les résidus semblent bien suivre une loi normale et être *iid* [1].

Nous pouvons retenir ce modèle.

1. indépendants et identiquement distribués

1.10 Représentation d'état

Un certain nombre de problèmes physiques dépendent d'états cachés, non observés. Ceci a conduit à développer en traitement du signal et en automatique les modèles dits d'espace d'état.

L'étude "classique" des séries chronologiques décompose le signal en une partie déterministe : tendance, saisonnalité et une partie aléatoire stationnaire. Un des inconvénients de cette décomposition est qu'elle est "rigide" en se basant sur la stationnarité qui limite son application. Un des intérêts du modèle d'état est d'assouplir cette décomposition en rendant la partie déterministe aléatoire et n'obligeant pas le signal à être stationnaire.

Un des outils les plus importants permettant de traiter l'étude d'un système d'état est le filtre de Kalman. Il permet de manière récursive aussi bien d'estimer que de prédire les variables d'état. Cependant, il est des fois difficile de mettre en œuvre ce filtre. En effet certains paramètres sont difficiles à identifier et l'on pourra remarquer que sa convergence dépend de l'initialisation de celui-ci.

1.10.1 Présentation des systèmes d'état

Un système d'état (II.24) est composé d'une :

- *équation d'observation* (1ère équation) : elle décrit comment les variables observées sont générées à partir des états cachés et des résidus.
- *équation d'état* (2nde équation) : elle décrit la manière dont les variables cachées sont créées. C'est là où réside la dynamique du système.

$$\begin{cases} y_t &= Z_t\alpha_t + \epsilon_t & \epsilon_t \sim N(0, H_t) \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t & \eta_t \sim N(0, Q_t) \end{cases} \tag{II.24}$$

Il est important de distinguer les variables observées (y_t) et celles qui ne le sont pas (α_t).

L'idée de base est que l'évolution du système au cours du temps est déterminée par le vecteur d'état α_t. Cependant comme α_t n'est pas observé directement, nous devons baser notre analyse sur y_t.

Les matrices $Z_t,\ T_t,\ R_t\ , H_t$ et Q_t sont dans un premier temps supposées connues et les erreurs $\epsilon_t,\ \eta_t$ sont supposées indépendantes.

La valeur du vecteur d'état initial α_1 est supposée suivre une loi normale $N(a_1, P_1)$ où α_1 est indépendante de $(\epsilon_1, \ldots, \epsilon_t)$ et $(\eta_1, \ldots, \eta_t)$. Dans un premier temps on peut supposer a_1 et P_1 connus.

Dans la pratique, il se peut que certains paramètres des matrices $Z_t,\ T_t,\ R_t\ , H_t$ et Q_t soient inconnus. On peut regrouper ces paramètres dans un vecteur ψ que l'on pourra estimer par maximum de vraisemblance.

Dans un grand nombre d'applications R_t est l'identité. Si ce n'est pas le cas on peut considérer le bruit $\eta_t^* = R_t\eta_t$ et $Q_t^* = R_tQ_tR_t'$. On se ramène à un modèle alors plus simple à étudier.

Cependant si R_t est de dimension $(m \times r)$ avec $r < m$ et Q_t n'est pas inversible, il est préférable de travailler avec η_t.

Dans la suite nous travaillerons avec $R_t = I$.

1.10.2 Lien des représentations d'état avec les modèles *VARMAX*

En physique on préférera exprimer les systèmes d'état sous la forme :

$$\begin{cases} y_t &= Ax_t + Bu_t + \epsilon_t \sim N(0, H_t) \\ x_{t+1} &= Cx_t + Du_t + \eta_t \sim N(0, Q_t) \end{cases} \tag{II.25}$$

avec y_t et u_t les vecteurs des données observées et x_t le vecteur d'état.

On peut remarquer qu'il y a équivalence avec le système précédent II.24 en posant :

$$\begin{cases} \alpha_t = \left(\begin{array}{c|c|c} x_t & \cdots & x_t \\ \hline \multicolumn{3}{c}{B^*} \\ \hline \multicolumn{3}{c}{D^*} \end{array}\right) \quad Z_t = \left(\begin{array}{c|c|c} & u_t^* & \\ A & \vdots & 0 \\ & u_t^* & \end{array}\right) \\ T_t = \left(\begin{array}{c|c|c} C & 0 & \begin{array}{c} u_t^* \\ \vdots \\ u_t^* \end{array} \\ \hline 0 & I & 0 \\ \hline 0 & 0 & I \end{array}\right) \quad R_t = \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{cases} \tag{II.26}$$

Ce modèle II.25 est équivalent aussi de la même manière à un processus *VARMAX*. En notant d l'opérateur de retard et z celui d'avance on obtient :

$$\begin{cases} (z - C)x_t &= Du_t + \eta_t \\ y_t &= Ax_t + Bu_t + \epsilon_t \end{cases} \tag{II.27}$$

on en déduit :

$$\begin{aligned} y_t &= A(z - C)^{-1}Du_t + Bu_t + A(z - C)^{-1}\eta_t + \epsilon_t \\ &= (Ad(I - Cd)^{-1}D + B)u_t + Ad(I - Cd)^{-1}\eta_t + \epsilon_t \end{aligned} \tag{II.28}$$

$$\begin{aligned} y_t &= A\sum_{k=0}^{\infty} C^k d^{k+1} Du_t + Bu_t + A\sum_{k=0}^{\infty} C^k d^{k+1}\eta_t + \epsilon_t \\ &= A\sum_{k=0}^{\infty} C^k Du_{t-k-1} + Bu_t + A\sum_{k=0}^{\infty} C^k \eta_{t-k-1} + \epsilon_t \end{aligned}$$

On peut remarquer qu'à la structure *VARMAX* correspond un grand nombre de modèles d'état. On pourra en privilégier physiquement certains et vouloir reconstituer X_t à partir de

Y_t et U_t observés. On peut étudier des représentations d'état de grandes dimensions en essayant de réduire cette dimension.

Les modèles d'état ne sont intéressants dans notre cas que si les bruits ε_t et η_t sont non nuls et laissent donc X_t le processus d'état caché. L'usage d'un filtre de Kalman se justifie dans un tel cas.

Le but de ce travail est alors d'estimer les matrices (A, B, C, D) contenant les paramètres physiques du bâtiment, ainsi que les matrices de covariance (H_t, Q_t). Nous avons pour cela utilisé l'algorithme EM [85] pour la partie initialisation du filtre puis suivant les procédures les plus efficaces l'algorithme de vraisemblance afin d'estimer les différentes matrices (A, C, H_t, Q_t) (annexe B.3), les matrices (B, D) étant estimées à partir du filtre de Kalman. Le détail de ces méthodes est présenté en annexe. Malgré un travail sur l'initialisation (Durbin and al. [35]) la qualité et le faible nombre de nos données d'apprentissage ne nous a pas permis de stabiliser les estimateurs. La cyclo-stationnarité semble aussi être un handicap pour appliquer le filtrage [33]. Une perspective sera de poursuivre ce travail afin de calculer la sensibilité par rapport aux paramètres ou de manière simultanée aux paramètres et aux entrées en appliquant la méthode développée dans le chapitre suivant.

Chapitre 2

Sensibilité pour des problèmes dynamiques

2.1 Définitions et estimations de la sensibilité dans le cas dynamique avec entrées dépendantes

2.1.1 Position du problème, $k-$sensibilité

Nous considérons dans la suite un modèle où les entrées dépendront du temps. La sortie sera elle aussi étudiée en prenant en compte la dimension temporelle.

Supposons que l'on ait un modèle entrée-sortie :

$$Y_t = \eta(\boldsymbol{U}_t, \ldots, \boldsymbol{U}_{t-k}), \quad t \in \mathbb{N}, \quad k \in \mathbb{N}, \quad \boldsymbol{U}_t \in \mathbb{R}^p$$

où k désigne la *mémoire* du système, k est indépendant de t.

La mémoire 0 correspond à un système à réaction *instantanée* : $Y_t = \eta(\boldsymbol{U}_t)$. La mémoire 1 correspond à $Y_t = \eta(\boldsymbol{U}_t, \boldsymbol{U}_{t-1})$. La mémoire sera dite totale si $Y_t = \eta(\boldsymbol{U}_t, \boldsymbol{U}_{t-1}, \ldots, \boldsymbol{U}_0)$, l'observation commençant à l'instant 0. Physiquement la mémoire de Y correspond par exemple à une inertie thermique dans les problèmes de chauffage liés aux propriétés des matériaux employés ou toute autre application où l'évolution de Y à un instant donné dépend de ses valeurs passées.
Dans la suite nous adopterons pour simplifier les notations suivantes :

- X désigne une variable aléatoire scalaire
- $\boldsymbol{X}$ désignera un vecteur aléatoire et $\boldsymbol{X}_t = (X_t^i)_{i=1,\ldots,p}$
- $\mathbb{X}_{t,k}$ désignera le k-vecteur $(X_t, X_{t-1}, \ldots, X_{t-k})$, si X_t est un processus stochastique scalaire ou vectoriel

La mémoire de Y est à distinguer de la mémoire du processus d'entrée $\boldsymbol{U}_t$. $\boldsymbol{U}_t$ est un processus multivarié se présentant le plus souvent sous la forme d'un processus $VAR(1)$. Ceci "cache" la mémoire puisque dans les coordonnées U_t^i de U_t peuvent intervenir des termes du type $U_t^j = U_{t-a}^i$. Le système Y reçoit une entrée ayant donc sa propre mémoire a et Y y ajoute sa

mémoire par le biais du modèle η. Par exemple si l'on considère le cas linéaire unidimensionnel :

$$Y_t = a_0\boldsymbol{U}_t + \cdots + a_k\boldsymbol{U}_{t-k} + \boldsymbol{\varepsilon}_t^Y \text{ avec } \boldsymbol{U}_t = b_0 + b_1\boldsymbol{U}_{t-1} + \cdots + b_k\boldsymbol{U}_{t-h} + \varepsilon_t^U$$

La mémoire effective du système entrée plus sortie par rapport aux innovations ε est de : $k+h$.

Le processus de sortie à l'instant t dépend de ses instants passés Y_{t-k} et par la même des instants passés du processus d'entrée $\boldsymbol{U}_{t-k}$. A cause du phénomène de mémoire, il n'est pas judicieux de calculer la sensibilité du système à l'instant t par rapport à l'entrée à l'instant t mais de calculer la sensibilité par rapport à $\mathbb{U}_{t,k}$. Ceci nous conduit à définir ce que nous appelons la $k-$*sensibilité* de Y_t par rapport à la composante U_t^1.

Pour simplifier l'exposé, essentiellement à cause de la lourdeur des notations dans le cas vectoriel, nous exposerons les résultats dans le cas où l'on cherche la sensibilité par rapport à une variable, alors on notera $\boldsymbol{Z}_t = (U_t^2, \ldots, U_t^p) \in \mathbb{R}^{p-1}$, $X_t = U_t^1$ et donc $\boldsymbol{U}_t = (X_t, \boldsymbol{Z}_t)$

Nous donnons les définitions suivantes :

Définition 7. *$k-$sensibilité*
la $k-$sensibilité *de Y par rapport à la composante X_t, $k > 0$ est définie par :*

$$S_{t,k}^X = \frac{\mathbf{Var}(\mathbf{E}(Y_t|\mathbb{X}_{t,k}))}{\mathbf{Var}(Y_t)}. \tag{II.1}$$

où $\mathbb{X}_{t,k}$ est le vecteur $(X_t, \ldots, X_{t-k})$

Il s'agit donc d'un indice mesuré par le rapport de la variance conditionnelle de Y_t lorsque $(X_t, \ldots, X_{t-k})$ est fixé à la variance totale de Y_t. La sensibilité instantanée correspond à $k = 0$.

Définition 8. *POPSI*
La sensibilité totale, c'est-à-dire par rapport à tout le passé de la composante X_t est donnée par :

$$S_t^X = \frac{\mathbf{Var}(\mathbf{E}(Y_t|\mathbb{X}_t))}{\mathbf{Var}(Y_t)}. \tag{II.2}$$

avec $\mathbb{X}_t = (X_t, \ldots, X_0)$.

Nous avons appelé S_t^X : Projection on the Past Sensitivity Index par rapport à X (indice POPSI) [52]. Le terme projection signifie simplement que l'opérateur $. \mapsto \mathbf{E}(.|\mathbb{X}_t)$ est une projection de l'espace des vecteurs aléatoires (de carré intégrable) sur l'espace des vecteurs fonction du processus $\mathbb{X}_t$.

On peut remarquer que la $\sigma-$algèbre des fonctions $f(X_t, \ldots, X_{t-k})$ croit lorsque k augmente. On a donc les propriétés suivantes :

Proposition 4.

$$\forall (t,k) \in \mathbb{N} \times \mathbb{N}, \quad S_{t,k}^X \leq 1 \tag{II.3}$$

$$\forall (t,k) \in \mathbb{N} \times \mathbb{N}, \quad S_{t,k-1}^X \leq S_{t,k}^X \tag{II.4}$$

$$t \geq 0 \ \max_{k \leq t} S_{t,k}^X = S_t^X \tag{II.5}$$

Pour un instant t fixé, lorsque la mémoire k croit, l'indice $S^X_{t,k}$ croit vers une valeur indépendante de k mais dépendante de t. En effet l'espace de projection, c'est à dire l'espace par lequel est conditionné la sortie, augmente au fur et à mesure que croit k : l'indice est donc croissant.

En pratique nous choisirons pour k la valeur à partir de laquelle l'indice $S^X_{t,k}$ ne croit plus significativement. Cette valeur heuristique k sera dite *mémoire utile* en terme de sensibilité. On peut la définir alors ainsi :

Définition 9. *Soit $\varepsilon > 0$. La mémoire utile est définie telle que :*

$$k_\varepsilon = inf\left\{k,\ \ |S^X_{t,h} - S^X_{t,k}| \leq \varepsilon,\ \ h > k\right\}$$

En général $\boldsymbol{Z}_t$ dépend de $\mathbb{X}_t$. Lorsque les entrées sont dépendantes la propriété $S^X_{t,k} \leq 1$ pour tout t et pour n'importe quelle mémoire k reste vraie. Cependant pour k et t fixés, la somme des indices est différente de 1 :

$$S^X_{t,k} + S^Z_{t,k} + S^{X,Z}_{t,k} \neq 1 \tag{II.6}$$

avec $S^{X,Z}_{t,k} = \dfrac{\mathbf{Var}\left(\mathbf{E}\left(Y_t|\mathbb{X}_{t,k}, \mathbb{Z}_{t,k}\right)\right)}{\mathbf{Var}Y_t} - S^X_{t,k} - S^Z_{t,k}$.

2.2 Extension de la méthode Pick and Freeze à des situations dynamiques et dépendantes

Supposons que l'on veuille estimer la sensibilité d'ordre k de la variable X_t par la méthode Pick and Freeze, il faut pouvoir réécrire le modèle sous la forme :

$$Y_t = g\left(\mathbb{X}_{t,k}, \mathbb{W}_{t,k}\right) \tag{II.7}$$

avec $\mathbb{W}_{t,k}$ un processus stochastique indépendant de $\mathbb{X}_{t,k}$.

Alors :

$$S^X_{t,k} = \frac{\mathbf{Var}\left(\mathbf{E}\left(g(\mathbb{X}_{t,k}, \mathbb{W}_{t,k})|\mathbb{X}_{t,k}\right)\right)}{\mathbf{Var}(Y_t)} \tag{II.8}$$

définit l'indice dans le cas indépendant où la méthode d'estimation Pick and Freeze s'applique.

On appelle copie de $(W_t)_{t\in\mathbb{N}}$ un processus $(W'_t)_{t\in\mathbb{N}}$ de mêmes lois marginales et indépendant de W_t. D'après le lemme de Sobol (1), si $(\mathbb{X}_{t,k}, \mathbb{W}_{t,k}, \mathbb{W}'_{t,k})$ sont trois processus vectoriels indépendants : $Y_t = g(\mathbb{X}_{t,k}, \mathbb{W}_{t,k})$ et $Y'_t = g(\mathbb{X}_{t,k}, \mathbb{W}'_{t,k})$ on a :

$$\mathbf{Var}(\mathbf{E}(Y_t|\mathbb{X}_{t,k})) = \mathbf{Cov}(Y_t, Y'_t)$$

On peut alors réécrire l'indice de sensibilité de la façon suivante :

$$S^X_{t,k} = \frac{\mathbf{Var}\left(\mathbf{E}(Y_t|\mathbb{X}_{t,k})\right)}{\mathbf{Var}(Y_t)} = \frac{\mathbf{Cov}(Y_t, Y'_t)}{\mathbf{Var}(Y_t)} \tag{II.9}$$

Estimer $S^X_{t,k}$ nécessite donc :

1. D'avoir une représentation du type $Y_t = g(\mathbb{X}_{t,k}, \mathbb{W}_{t,k})$
2. De pouvoir simuler un $n-$échantillon de $\mathbb{W}_{t,k}$

La loi des grands nombres appliquée aux processus $(\mathbb{X}_{t,k}, \mathbb{W}_{t,k}, \mathbb{W}'_{t,k})$ permet d'obtenir un estimateur de l'indice de sensibilité.

2.2.1 Cas Gaussien

Soit $\boldsymbol{U}_t$ un processus vectoriel gaussien centré donné et un système entrée-sortie donné par :

$$\begin{aligned} Y_t &= \eta(\boldsymbol{U}_t, \boldsymbol{U}_{t-k}, \dots, \boldsymbol{U}_0) \\ &= \eta(X_t, X_{t-1}, \dots, X_0, \boldsymbol{Z}_t, \boldsymbol{Z}_{t-k}, \dots, \boldsymbol{Z}_0) \end{aligned}$$

On cherche la sensibilité de Y_t par rapport à $\mathbb{X}_{t,k} = (X_t, X_{t-1}, \dots, X_{t-k})$.

Afin d'appliquer la méthode Pick and Freeze, il nous faut remplacer toutes les variables différentes de $(X_t, X_{t-1}, \dots, X_{t-k})$ par leurs décompositions par rapport à $\mathbb{X}_{t,k}$.

Soit $A_{t,k} = \{\mathbb{X}_{0,t-k-1}, \mathbb{Z}_t\}$. Pour tous $\phi \in A_{t,k}$, ϕ peut être décomposée en deux parties orthogonales indépendantes :

$$\phi = \tilde{\phi}_{t,k} + W^{\phi}_{\mathbb{X}_{t,k}} \tag{II.10}$$

ou si il n'y a pas de confusion possible :

$$\phi = \tilde{\phi}_{t,k} + W^{\phi}_{t,k} \tag{II.11}$$

avec $\tilde{\phi}_{t,k}$ la projection orthogonale de ϕ sur l'espace engendré par les variables aléatoires $(X_t, \dots, X_{t-k})$.

On peut ainsi réécrire le modèle entrée-sortie par :

$$Y_t = g(\mathbb{X}_{t,k}, \mathbb{W}_{t,k}) \tag{II.12}$$

où $\mathbb{W}_{t,k}$ est :

$$\mathbb{W}_{t,k} = \left(W^{\mathbb{X}_{0,t-k-1}}_{t,k} \ , W^{\mathbb{Z}_t}_{t,k}\right) \tag{II.13}$$

Dans le cas Gaussien nous sommes exactement dans la situation où les entrées $\mathbb{X}_{t,k}$ et $\mathbb{W}_{t,k}$ sont indépendantes.

On observe que $\tilde{\phi}_{t,k}$ est une fonction linéaire de $\mathbb{X}_{t,k}$ alors :

$$\begin{aligned} \tilde{\phi}^{t,k} &= \sum_{u=t-k}^{t} \lambda^{\phi}_u X_u \\ &= (\boldsymbol{\lambda^{\phi}})^* \mathbb{X}_{t,k} \ , \quad \text{avec } \boldsymbol{\lambda^{\phi}} = (\lambda^{\phi}_u)_{t-k \le u \le t} \end{aligned}$$

On obtient $\boldsymbol{\lambda^{\phi}}$ à partir de :

$$\hat{\boldsymbol{\lambda}}^{\phi} = \underset{\mu}{argmin} \left(\mathbf{E}|\phi - \boldsymbol{\mu}\mathbb{X}_{t,k}|^2\right) \tag{II.14}$$

Alors si la matrice bloc de covariance $\Gamma^X_{t,k} = \mathbf{E}(\mathbb{X}_{t,k}\mathbb{X}^*_{t,k})$ du processus X de dimension $(k+1)\times(k+1)$ est inversible, un estimateur de $\boldsymbol{\lambda}^\phi$ est :

$$\hat{\boldsymbol{\lambda}}^\phi = \left(\Gamma^X_{t,k}\right)^{-1}\gamma^{X,\phi}_{t,k} \tag{II.15}$$

où $\gamma^{X,\phi} = \mathbf{E}(\mathbb{X}_{t,k}\phi)$ est le vecteur de covariance de taille $(k+1)$ entre les processus $\mathbb{X}_{t,k}$ et ϕ.

$\mathbb{W}_{t,k}$ est obtenu par :

$$\mathbb{W}_{t,k} = \phi - \tilde{\phi}^{t,k} \tag{II.16}$$

Pour appliquer la méthode Pick and Freeze pour un indice de mémoire k, on doit obtenir par simulation N couples $(X^i_{s,t}, \mathbb{W}^i_t)$, $(X^i_{s,t}, \mathbb{W}'^i_t)$ avec $\mathbb{W}_t$ et $\mathbb{W}'_t$ indépendants. Nous avons supposé a priori que $\mathbb{U}_t$ est assez aisément simulable. Les différentes étapes pour calculer les indices de Sobol sont :

1. Simuler deux échantillons de tailles N : $(\boldsymbol{U}_t = (X_t, \boldsymbol{Z}_t))_{t\in\mathbb{N}}$ et $(\boldsymbol{U}'_t = (X'_t, \boldsymbol{Z}'_t))_{t\in\mathbb{N}}$
2. Pour tout $\phi \in A_{t,k}$ et $\phi' \in A'_{t,k}$, calculer :

$$\begin{aligned}\boldsymbol{\lambda}^\phi_{t,k} &= \left(\Gamma^X_{t,k}\right)^{-1}\gamma^{X,\phi}_{t,k}\\ \tilde{\phi}_{t,k} &= \boldsymbol{\lambda}^\phi_{t,k}\mathbb{X}_{t,k}\\ \tilde{\phi}'_{t,k} &= \boldsymbol{\lambda}^\phi_{t,k}\mathbb{X}'_{t,k}\end{aligned}$$

   $A_{t,k}$ joue le rôle de $A'_{t,k}$ quand $\mathbb{U}'_{t,k}$ joue le rôle de $\mathbb{U}_{t,k}$
3. Calculer :

$$\begin{aligned}\mathbb{W}_{t,k} &= \phi - \tilde{\phi}^j_{t,k}, \quad \phi \in A_{t,k}\\ \mathbb{W}'_{t,k} &= \phi' - \tilde{\phi}'^j_{t,k}, \quad \phi' \in A'_{t,k}\end{aligned}$$

4. Créer un autre échantillon où la partie correspondante à $\mathbb{X}_{t,t}$ est gelée :

$$\boldsymbol{U}^X_t = (\mathbb{X}_{t,k}, \mathbb{W}'_{t,k} + \tilde{\phi}_{t,k})$$

5. Calculer (Y, Y^X) à partir des processus $(\boldsymbol{U}_t, \boldsymbol{U}^X_t)$
6. Calculer l'indice à partir de la formule I.29 ou I.30

On pourra trouver le pseudo-code en partie appliquée 5.1.

Exemples :

Voici des exemples d'applications de la méthode de séparation des variables. Considérons deux modèles jouet : un linéaire et un qui ne l'est pas donnés par :

$$Y_t = 0.2Y_{t-1} + 0.3X_t + Z_t \tag{II.17}$$
$$Y_t = X_tZ_t + 0.2exp(-Z_t) \tag{II.18}$$

X_t, Z_t sont des processus $VAR(1)$ stationnaires tels que :

$$\begin{pmatrix} X_t \\ Z_t \end{pmatrix} = \begin{pmatrix} 0.8 & 0.4 \\ 0.1 & 0.2 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Z_{t-1} \end{pmatrix} + \omega_t \tag{II.19}$$

où ω_t est un bruit stationnaire gaussien de matrice de covariance : $\Theta = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$.

Tous les programmes ont été réalisés sous R.

Présentons d'abord les résultats pouvant être obtenu par l'algorithme 3 lorsque la sensibilité de Y par rapport à X est calculé pour le modèle (II.17). Λ est le vecteur $(\lambda_t, \lambda_{t-1}, \ldots, \lambda_{t-k}, \ldots, \lambda_0)^*$. Ces valeurs sont données dans le tableau : 2.1. Après avoir simulé $Simu_1$ (table : 2.2), $\tilde{X}_1$ est calculé. A partir des valeurs de Λ données dans la table 2.1 on calcul les valeurs correspondant à $\tilde{X}$ définit comme précédemment. La table 2.2 est le résultat de l'étape (17) de 3 : $W_1 = Z - \tilde{X}_1$. On peut faire le même travail pour $Simu_2$ et obtenir W_2. On peut remarquer que les coefficients de Λ décroissent. Seul les trois premier instant passé sont important (table 2.1).

time \ time	0		1		2		3		4	
0	X_t	0.12	X_{t-1}	0.38	X_{t-2}	0.07	X_{t-3}	-0.01	X_{t-4}	-0.01
1			X_t	-0.21	X_{t-1}	0.33	X_{t-2}	0.07	X_{t-3}	0.00
2					X_{t-1}	-0.21	X_{t-1}	0.33	X_{t-2}	0.07
3							X_t	-0.21	X_{t-1}	0.33
4									X_t	-0.21

TABLE 2.1 – Valeurs estimées de Λ (Etape (13) de l'algorithme 3)

time	X_t	Z_t	$\tilde{X}_t$	W_t
0	-0.21	0.40	-0.02	0.42
1	0.11	-0.12	-0.10	-0.02
2	-0.17	- 1.02	0.06	-1.08
3	-0.29	-0.79	0.01	-0.80
4	0.24	-0.80	-0.16	-0.64

TABLE 2.2 – Valeurs de X, Z, $\tilde{X}_t$ et W pour $Simu_1$ (Etape (8) de l'algorithme 3)

Remarque 8. *Si le processus est stationnaire la covariance est invariante par translation dans le temps. Alors :* $S_{t,k} = S_k$

Sur chaque figure II.1, II.2, II.3, II.4, les indices $S_t^X = \dfrac{\mathbf{Var}(\mathbf{E}(Y_t|\mathbb{X}_t))}{\mathbf{Var}(Y_t)}$ sont tracés à chaque pas de temps, calculé pour les modèles (II.17) et (II.18) pour des échantillons de différentes tailles ($N = 200$ et $N = 10000$). Les intervalles de confiances à 95% sont dessinés sur chaque figure. Ces exemples exhibent deux types de convergence :

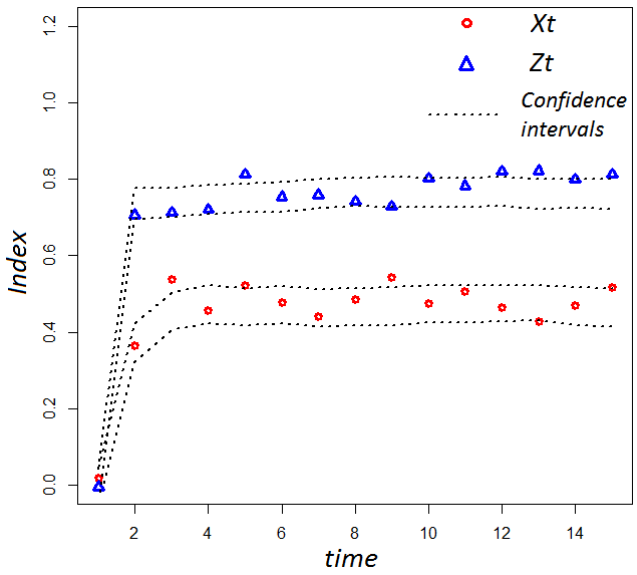

FIGURE II.1 – Modèle jouet II.17 : Indices estimés de Sobol en fonction du temps. Echantillon de taille : 200. Intervalle de confiances à 95% en pointillé.

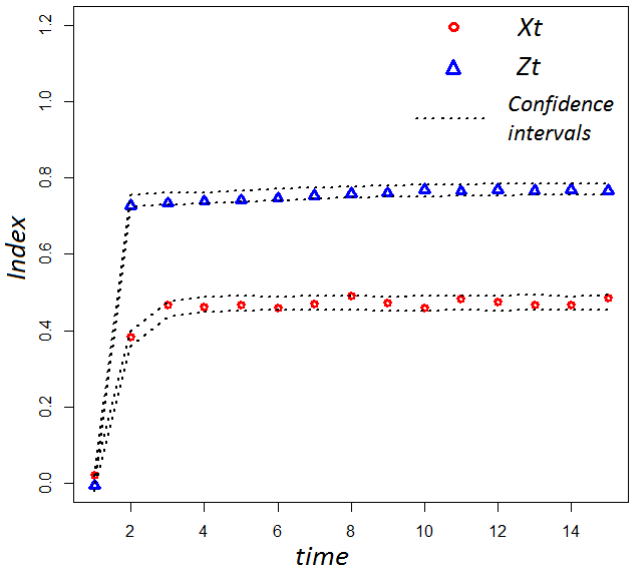

FIGURE II.2 – Modèle jouet II.17 : Indices estimés de Sobol en fonction du temps. Echantillon de taille : 10000. Intervalle de confiances à 95% en pointillé.

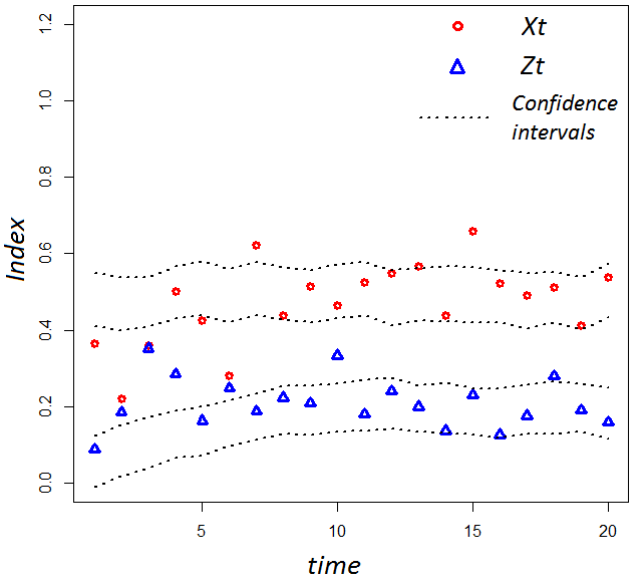

FIGURE II.3 – Modèle jouet II.18 : Indices estimés de Sobol en fonction du temps. Echantillon de taille : 200. Intervalle de confiances à 95% en pointillé.

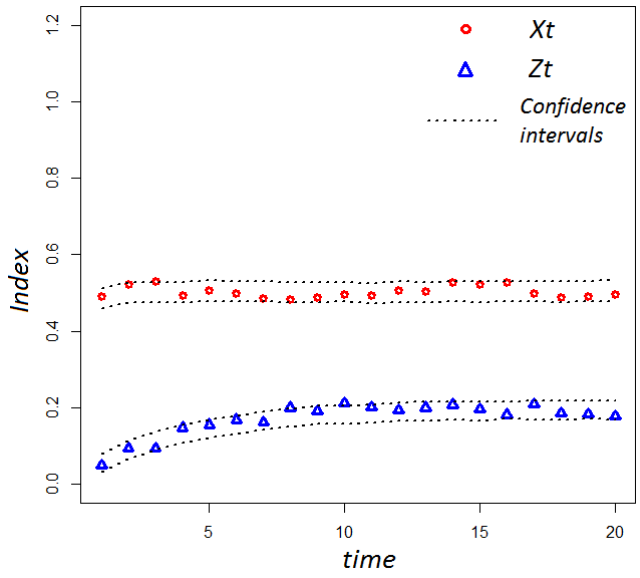

FIGURE II.4 – Modèle jouet II.18 : Indices estimés de Sobol en fonction du temps. Echantillon de taille : 10000. Intervalle de confiances à 95% en pointillé.

— la convergence de l'estimateur. A chaque pas de temps, l'algorithme Pick and Freeze estime S_t^X. La qualité de l'estimateur $\widehat{S}_t^X$ dépend de la taille de l'échantillon utilisée. L'intervalle de confiance diminue lorsque $N = 10000$. La vitesse de convergence de l'estimateur est lente ($O(1/\sqrt{N})$) lorsque l'on utilise une méthode de type Monte Carlo. On pourrait améliorer cette vitesse en utilisant une méthode de type Quasi Monte Carlo (QMC). Dans notre cas cette méthode nous a semblé difficile à implémenter.

— une convergence temporelle. L'indice $\widehat{S}_t^X$ change au cours du temps. Après quatre itérations, l'indice converge vers une constante comme par exemple sur les figures : II.1, II.2. Le modèle II.17 est auto-régressif, ce qui signifie que Y_t dépend de l'instant passés Y_{t-1} et donc, de tous les instants passé des entrées X_t, Z_t. On peut réécrire le II.17 par :

$$Y_t = \sum_{k=0}^{\infty} (0.2)^k (0.3X_{t-k} + Z_{t-k})$$

A l'instant $t = 0$ par exemple, $S_0^X = \dfrac{\mathbf{Var}(\mathbf{E}(Y_0|X_0))}{\mathbf{Var}(Y_0)}$. Y_0 est projeté sur un espace de dimension 1 alors qu'il dépend de $(X_{-\infty}, \ldots, X_0, Z_{-\infty}, \ldots, Z_0)$. L'espace de projection est trop petit. En agrandissant cet espace, l'indice augmente et converge vers une constante. C'est pourquoi nous avons défini les phénomènes de mémoire qui peut référer au concept physique d'inertie. Lorsque Y_t dépend uniquement de (X_t, Z_t) l'indice converge directement (figures : II.3 et II.4).
La convergence temporelle, lorsque les entrées sont stationnaires, est intéressante du point de vue computationel. La convergence de l'indice vers une constante, rend inutile de calculer l'indice sur toute sa trajectoire mais seulement sur une portion fortement réduite (jusqu'à la convergence). Ceci représente une économie de calcul et donc de temps.

On peut aussi noter que la méthode développée est indépendante du modèle utilisé f. C'est une méthode boîte noire, nécessitant seulement la possibilité de simuler un nombre important d'entrées et par la même de sorties du modèle.

2.2.2 Méta-modèles dynamiques non gaussiens et sensibilité

La construction de méta-modèles non gaussiens de processus de structure donnée, comme par exemple les processus $VAR(1)$ est un problème complexe en devenir du point de vue mathématique. Il est clair pour les problèmes dynamiques que l'on ne peut pas poser les problèmes de sensibilité par rapport à un paramètre physique comme des problèmes statiques (du moins en général). L'exemple gaussien montre bien la spécificité de la sensibilité pour une variable aléatoire évoluant dans le temps.

Les lois de bien des variables ne sont cependant pas gaussiennes. Ainsi des variables climatiques (température, vent) sont le plus souvent bornées (on le voit en appliquant la théorie des extrêmes). Il en est de même des variables de type chauffage ou source d'énergie. Le phénomène peut ne pas être grave si l'on se désintéresse des valeurs extrêmes comme dans la littérature concernant le bâtiment. Si l'on veut étudier l'impact d'un froid extrême et encore plus d'une vague de chaleur sur la température intérieure, on ne peut utiliser des variables gaussiennes

à queues trop lourdes. Il en est de même pour des phénomènes présentant par exemple deux valeurs principales pour le facteur étudié avec une transition aléatoire plus ou moins régulière. La loi de la variable, dans ce cas, n'est pas gaussienne mais bi-modale.
Supposons pour se fixer les idées, vouloir déterminer l'existence et estimer les paramètres d'un processus $VAR(1)$ soumis aux contraintes suivantes :

— Les lois marginales d'ordre 1 des p variables $(X^1, \dots, X^p)$ sont fixées
— La covariance Γ (ou bien la dynamique A) de l'équation : $\boldsymbol{X}_t = A\boldsymbol{X}_{t-1} + \varepsilon_t$, est fixée

Les deux contraintes ne sont pas de même nature sauf si l'ensemble d'apprentissage est très grand. Estimer ou choisir p lois demande beaucoup d'observations. Travaillant sur des méta-modèles, les propriétés qualitatives sont les plus importantes : le support, la forme des queues de probabilité, le nombre de modes sont les trois qualités que l'on envisage souvent.
Estimer la covariance se fait de façon empirique et nécessite moins de données. De plus si l'on se fixe le modèle, la vraisemblance gaussienne est toujours une fonction de contraste ([26]) qui donne des estimateurs convergents. La relation $(I - A)\Gamma^{X_0} = \Omega_\varepsilon$ permet de connaitre A à partir de la covariance de $\boldsymbol{X}_0$ en régime stationnaire et de la covariance des bruits, quantités estimables à partir du contraste gaussien. En particulier si $\Omega_\varepsilon = I$ on a directement : $A = (\Gamma^{X_0})^{-1} - I$.

Nous appellerons *problème* VAR(1) *non gaussien* le problème suivant :

— Les fonctions de répartition $F_1, \dots, F_p$ des marginales de X_t, X_t processus stationnaire sont données par des densités $f_1, \dots, f_p$ fixées (et donc indépendantes de t).
— la covariance $\Gamma = \left(\mathbf{E}(X^i_\alpha X^j_\beta)\right)_{\substack{\alpha,\beta\in[t,(t-1)]\\ i,j=1,\dots,p}}$ est donnée ou A et Ω_ε sont données.

Nous traitons le problème dans le cas où :

$$\boldsymbol{X}_t = A\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t$$

Remarquons que dans ce cas $\mathbf{Cov}(X_t, X_{t-1})$ est connue à partir des équations :

$$\begin{aligned}
&\mathbf{E}(\boldsymbol{X}_t \boldsymbol{X}^*_{t-1}) = A\Gamma^X \\
&\Gamma^X = A\mathbf{E}(\boldsymbol{X}_{t-1}\boldsymbol{X}^*_t) + \mathbf{E}(\boldsymbol{\varepsilon}_t \boldsymbol{X}^*_t) \\
&\mathbf{E}(\boldsymbol{\varepsilon}_t \boldsymbol{X}^*_t) = \mathbf{E}(\varepsilon_t \boldsymbol{\varepsilon}^*_t)
\end{aligned}$$

et donc par la suite toutes les covariances $\mathbf{Cov}(\boldsymbol{X}_t, \boldsymbol{X}_{t-h})$.

Définition 10. *Un problème $(F_1, \dots, F_p, \Gamma)$ est dit réalisable s'il admet au moins une solution.*

La stationnarité simplifie évidemment le problème puisque la donnée de Γ^{X_0} et Γ^ε suffit à déterminer toutes les structures du second ordre, en particulier A.

Nous ne détaillons pas les contres exemples mais il existe pour tout $p \geq 2$ des modèles non réalisables. En fait et cela est évident sur des exemples simples, la donnée de $F_1, \dots, F_p$ peut être incompatible avec la donnée de Γ.

Dans la suite pour simplifier (cela n'est pas essentiel) nous supposons $\Omega_\varepsilon = I$. Il ne semble pas que dans les études de sensibilité le problème soit posé en termes de distribution du bruit.

Ayant un processus $VAR(1)$ stationnaire de covariance connue et souhaitant avoir des lois marginales non gaussiennes (par exemple bornées ou bimodales) dans la modélisation, comment peut-on essayer de réaliser de telle loi réaliser ?

Par exemple pour des lois uniformes, on sait qu'en 2 dimensions, on ne peut pas les réaliser par des copules gaussiennes pour toute valeur de coefficient de corrélation mais on peut les réaliser pour d'autres copules. De nombreux travaux ont été faits mais les plus marquants sont fondés sur les lois marginales introduites par Johnson [71]. Elles constituent un ensemble de lois paramétriques qui permettent en dimension 2 de choisir un modèle dont les marginales ont une moyenne, une variance, une skewness (asymétrie mesurée par le moment d'ordre 2), une kurtosis (première mesure du poids de la queue de probabilité) fixées et surtout un coefficient de corrélation donné, soit cinq paramètres.

Le système de lois de Johnson est défini ainsi :

La fonction de répartition s'écrit :

$$F_X(x) = \Phi(\gamma + \delta f|(x-\xi)/\lambda|) \tag{II.20}$$

où γ et δ sont les paramètres de forme, ξ paramètre de position, λ paramètre d'échelle et $f(\dot{})$ est l'une des transformations suivantes :

$$f(y) = \begin{cases} \log(y) & \text{famille lognormale} \\ \log(y + \sqrt{y^2+1}) & \text{lois non bornées} \\ \log(\dfrac{y}{1-y}) & \text{lois bornées type logistique} \\ y & \text{lois normales} \end{cases} \tag{II.21}$$

Pour chaque famille, moyenne, variance, skewness et kurtosis déterminent $(\gamma, \delta, \xi, \lambda)$ et réciproquement. Le nombre maximal de modes est 2.

Un processus $VAR(1)$ de Johnson est un méta-modèle dont les marginales $F_1, \ldots, F_p$ sont des répartitions de Johnson et dont la corrélation si le processus est normé (ou la covariance sinon) est donnée par une matrice de type positif :

$$\Gamma = (r_{i,j})_{i=1,\ldots,p;j=1,\ldots,p}$$

Le processus gaussien à l'origine du processus de Johnson est un processus $\boldsymbol{Z}_t$ de lois $\mathcal{N}(0, \Gamma^Z)$.

On a donc :

$$\begin{aligned} X_t^i &= \overleftarrow{F}_i(\Phi(Z_t^i)) \\ r_{i,j} &= \mathbf{Cov}(\overleftarrow{F}_i(\Phi(Z_t^i)), \overleftarrow{F}_j(\Phi(Z_t^j))) \end{aligned}$$

qui se calcule par une intégrale double à partir de la loi gaussienne de (Z_t^i, Z_t^j).
On posera :

$$\begin{aligned} \rho_{i,j} &= \mathbf{E}(Z_t^i Z_t^j) \\ \Gamma^Z &= (\rho_{i,j})_{i=1,\ldots,p;j=1,\ldots,p} \end{aligned}$$

Enfin, nous avons besoin de l'expression de X_t^i en fonction de Z_t^i soit :

$$X_t^i = \xi_t + \lambda \overleftarrow{f}_i(\frac{Z_t^i - X_t}{\delta_i})$$

On voit que $\overleftarrow{f}_i$ étant monotone, X_t^i est une fonction monotone de Z_t^i la fonction réciproque étant partout définie, cette correspondance bi-univoque montre que :

$$\sigma(X_t^i) = \sigma(Z_t^i)$$

On pose alors : $r_{i,j} = r_{i,j}(\rho_{i,j})$

Suivant des résultats de Biller et Nelsen [88] on a :

Théorème 4. *$c_{i,j}$ est une application continue de $[0,1]$ dans $[-1,1]$, qui conserve le signe, $c_{i,j}(0) = 0$ et $c_{i,j}(-1) = \rho_{i,j}$, $c_{i,j}(1) = \bar{\rho_{i,j}}$, où ρ et $\bar{\rho}$ sont les conditions minimales et maximales réalisables pour le couple (F_i, F_j) $c_{i,j}$ est monotone croissante.*
Enfin $c_{i,j}(\rho) = 0$ implique non seulement $r = 0$ mais aussi l'indépendance de X_t^i et X_t^j.

Théorème 5. *$(c_{i,j}(\rho_{i,j}))_{i=1,\dots,p;\ j=1,\dots,p}$ est une matrice de type positif.*

Toute matrice positive ne peut être obtenue de cette manière. On ne peut pas toujours réaliser de cette manière (dite $VARTA$) tout système $(F_1, \dots, F_p, \Gamma^X)$ ce qui est une limite sérieuse. En général lorsque la dimension p augmente, les matrices Γ^Z obtenues à partir de Γ^X sont de plus en plus fréquemment de type "non positif". Cela se voit par simulation. Biller et al.[8] fournissent un logiciel permettant de calculer $c_{i,j}$ et donc de calculer $\rho_{i,j}$ à partir de données de $r_{i,j}$.

Comme il s'agit de construire un méta-modèle qui est en partie approximatif (les lois de Johnson n'ont pas de réalité physique), on peut approcher la matrice $\Gamma^Z = \left(\overleftarrow{c}_{i,j}(r_{i,j})\right)_{i,j=1,\dots,p}$ par une matrice de type positif.

Soit $Y = \eta(X^1, \dots, X^p)$ un système entrée-sortie où $X^1, \dots, X^p$ ont des lois $F_1, \dots, F_p$ de Johnson. En remplaçant X^j par son expression en fonction de V^j variable normale on a :

$$Y = \tilde{\eta}(V^1, \dots, V^p)$$

qui assure que les σ-algèbres engendrées par X^j est V^j sont identiques. Les opérateurs $\mathbf{E}^{X^j}$ et $\mathbf{E}^{V^j}$ sont égaux, soit :

$$\mathbf{E}(\eta(X^1, \dots, X^p)|X^j) = \mathbf{E}(\tilde{\eta}(V^1, \dots, V^p)|V^j) \tag{II.22}$$

Prenons $j = 1$ pour simplifier les notations. Nous sommes dans le cas d'un problème réalisable (ou quasi réalisable si l'on utilise le résultat de Ghosh et al. [49]) donc $(V^1, \dots, V^p)$ a une matrice Γ_1 de covariance déduite de Γ (programmée par Biller et al. [8]).

Nous sommes ainsi ramenés au cas gaussien pour calculer la sensibilité par la méthode Pick and Freeze.

Remarque 9. *Les lois de Johnson permettent de passer d'un système $(F_1, \dots, F_p, \Gamma)$ à une copule gaussienne $\mathcal{N}(0, \Gamma_1)$. Cette correspondance permet de calculer la sensibilité par rapport à un facteur. Nous pensons que ce même résultat vaut pour tout groupe de facteurs. Pour d'autres lois que celle de Johnson la fonction c_y n'a pas a priori des propriétés assez bonnes pour expliciter Γ_1. En particulier on ne sait pas si la relation $c_{i,j}(\rho) = 0$ implique l'indépendance.*

En résumé le système $(F_1, \ldots, F_p, \Gamma)$ permet de déterminer Γ_0 de façon univoque et si $\sigma(X^1) = \sigma(V^1)$ alors on pourra estimer par exemple par Pick and Freeze la sensibilité en se ramenant au cas Gaussien.

Une fois cette remarque faite on sait calculer par Pick and Freeze les sensibilités dans le cas gaussien et donc le calcul se transfère au cas de modèles de Johnson. Le même résultat peut être transféré aux sensibilités d'ordre 2, nous ne savons pas si l'on peut aller au delà, à savoir si $\sigma(X^1, \ldots, X^k) = \sigma(V^1, \ldots, V^k)$ pour $k \geq 3$.

Programme pour construire un modèle non gaussien et calculer la sensibilité d'ordre 1 par Pick and Freeze

1. Calculer empiriquement les 4 premiers moments de $X^i, i = 1, \ldots, p$
2. Estimer (par exemple par la méthode des noyaux) la densité de probabilité g de X^1. Eventuellement estimer le nombre de modes et le support.
   En déduire la forme de la transformation f intervenant dans les lois de Johnson.
3. Estimer empiriquement en utilisant le contraste gaussien la covariance Γ^X et A ou A et Ω_ε. En déduire l'équation : $\boldsymbol{X}_t = A\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t$.
4. Utiliser le logiciel de Biller et Johnson pour calculer la matrice Γ^Z. Si Γ^Z n'est pas positive, l'approcher par une matrice de type positif.
5. Appliquer la méthode Pick and Freeze au processus gaussien de coordonnées $(V^1, \ldots, V^p)$ de covariance Γ pour calculer $\mathbf{E}^{V^1}$ puis calculer la sensibilité $S^{V^1} = S^{X^1}$.

Cependant Gosh et al. [49] ont donné des résultats très intéressants en pratique. Si Γ_1 est l'image de Γ, si l'on met sur les matrices de type positif la mesure uniforme alors il existe au voisinage de Γ_1 une matrice de type positif ce qui semble suffisant pour construire de manière raisonnable un méta-modèle. Ces résultats sont à confirmer.

2.3 Sensibilité, données brutes et données réduites

La modélisation avec pour fin en particulier la simulation exige le plus souvent de passer à des données réduites. Dans le cas d'applications pratiques, il est rare que le système d'entrées soit stationnaire. Il est souvent possible de se ramener à un tel cas en centrant et normalisant les processus $\boldsymbol{U}_t$:

$$\boldsymbol{U}_t = \mathbf{E}(\boldsymbol{U}_t) + \boldsymbol{s}_t \boldsymbol{V}_t$$

Étudions d'abord comment se comporte l'indice si les entrées sont réduites. Remarquons d'abord que l'indice de Sobol est le quotient de deux variances de la sortie Y, la variance usuelle et la variance de l'espérance conditionnelle. Cet indice est invariant par addition à Y d'une constante. Donc si la relation :

$$Y_t = \eta(X_t, X_{t-1}, \ldots)$$

est linéaire alors la sensibilité est invariante par centrage du processus d'entrée. Si la relation entrée-sortie n'est pas linéaire on ne peut en général rien dire.

Pour la variance même dans le cas linéaire, la situation n'est pas simple. En effet la variance est calculée pour chacune des variables d'entrée et donc il n'y a pas d'homogénéité.

Un cas particulier à ce sujet est très important, c'est celui des processus d'entrée cyclo-stationnaires. La covariance est une fonction matricielle périodique dont les variances sont la diagonale. Il n'y a pas lieu de normer les variables puisque le formalisme va tenir compte de la périodicité des variances.

Le schéma général pour pouvoir réaliser une analyse de sensibilité est donné sur la figure : II.5.

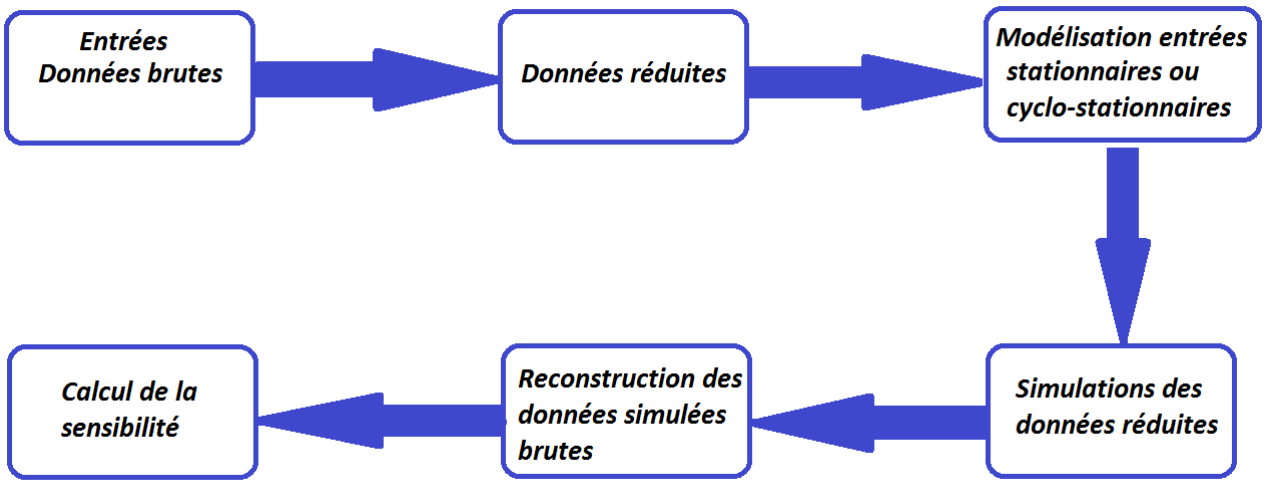

FIGURE II.5 – Schéma général de génération de données nécessaires à une analyse de sensibilité

2.4 Tracé des indices

Les indices calculés dépendent en général de t l'instant auquel il est calculé et de la taille de la mémoire k que l'on souhaite. Par exemple si l'on veut calculer l'indice de mémoire 1, il faut se placer à cet instant et donc faire t fois le calcul de l'indice. Si l'on veut tracer toutes les mémoires il faudra alors $k * t$ calculs. L'avantage de travailler avec une sortie stationnaire ou cyclo-stationnaire est de réduire considérablement le nombre de calculs. En effet si le processus est stationnaire la covariance est invariante par translation dans le temps. Si l'on calcule alors l'indice de mémoire 1, par exemple, à l'instant t_1, on obtient les indices de mémoire 1 pour tous les instants t. On a donc :

$$S_{k,t} = S_k \tag{II.23}$$

On gagne alors un nombre considérable de calculs. Si l'on veut tracer toutes les mémoires il faudra alors k calculs.

Si la sortie est cyclo-stationnaire de périodes P, les covariances sont périodiques de périodes

P. Ainsi pour k fixé, on a :

$$S_{k,t} = S_{k,t+hP} \quad h \in \mathbb{Z} \tag{II.24}$$

Pour obtenir tous les indices il faut alors calculer $S_{t,k}$ pour $t \in [0, P]$, soit $P * k$ calculs.

Il est donc préférable de se ramener à des cas stationnaires ou cyclo-stationnaires afin de réduire considérablement les calculs. Pour se ramener à un cas stationnaire il suffit souvent de centrer et normaliser les processus. Nous avons vu précédemment que l'indice de sensibilité est invariant par normalisation et centrage de la sortie. Se ramener à un cas stationnaire présente l'avantage donc de réduire considérablement le temps de calcul sans changer la valeur de l'indice.

2.5 Sensibilité par rapport à un groupe de variables

Il est aisé de voir qu'il n'y a que peu de choses à changer par rapport au cas unidimensionnel si ce n'est les notations et le formalisme des calculs.

Par exemple les matrices $\Gamma_{k,t}^{XX}$ et $\gamma_{k,t}^{X\phi}$, lorsque X est un vecteur, doivent être entendues comme des matrices se calculant par blocs. Ainsi $\Gamma_{u,v}^{XX}$ est le bloc de terme général $(\mathbf{E}(X_u^j X_v^l))$ pour $1 \leq j, l \leq J, \;\; k \leq u, v \leq t$.
On a évidement :

$$S_{t,k}^J \geq S_{t,k}^{J'} \text{ si } J > J' \tag{II.25}$$

d'après les propriétés de l'espérance conditionnelle et ce dans tous les cas si $S_{t,k}^J$ désigne la k-sensibilité par rapport au groupe de variables aléatoires $(U_v^j, \;\; j \in J, \;\; t - k \leq v \leq t)$

Chapitre 3

Conclusion

Ayant privilégié la méthode Pick and Freeze et souffrant d'un manque important de données pour notre application nous avons rappelé les méthodes utilisées pour construire des méta-modèles dynamiques et leurs propriétés. Ces modèles seront utilisés dans la partie suivante pour modéliser les différentes variables.

Nous avons adapté la définition de l'indice de Sobol à un cadre dynamique. Nous conditionnons la variance de l'espérance de la sortie à un instant t donné par rapport aux valeurs passées d'une variable d'entrée. Cet indice varie suivant les instants considérés dans le passé. Les propriétés de ce nouvel indice montrent que pour un instant t lorsque l'on augmente le passé de la variable d'entrée considéré, l'indice tend vers une constante. On a donc défini ce que l'on a appelé : mémoire utile. La mémoire utile est la longueur du passé nécessaire pour que l'indice n'augmente plus. Cette notion renvoie à des notions d'inertie dans un cadre physique.

Dans le cas Gaussien, nous avons développé une méthode permettant de calculer ces indices à partir de la méthode Pick and Freeze. Les variables d'entrée sont séparées en une partie dépendante des variables par lesquelles on conditionne la sortie et une partie indépendante de celle-ci. Sur ce nouveau jeu de variables indépendantes on applique la méthode Pick and Freeze pour calculer l'indice. Le pseudo-code de cette méthode est donné dans la partie suivante ainsi que les résultats obtenus pour ces indices.

Dans le cas non Gaussien une voie de travail est donnée. Pour des processus stationnaires si les p marginales ne sont pas gaussiennes, par exemple bimodales, en les choisissant par exemple dans une famille paramétrique (par exemple celle de Johnson), il est possible pour des processus de type VAR de donner un métamodèle satisfaisant à la fois les p contraintes de marginales et la contrainte de dynamique en utilisant une $p-$copule gaussienne. Sous certaines conditions on a démontré qu'il est possible de calculer les indices de Sobol du premier ordre uniquement à partir de processus gaussiens originaires de la copule. Sinon, il existe des moyens d'approcher la matrice de copule par une matrice de type positif mais cela reste à affiner.

Troisième partie

Application à un problème de bâtiment

Chapitre 1

Enjeux de l'analyse de sensibilité en énergétique des bâtiments

En 2008, les 27 pays de l'Union Européenne se sont engagés en adoptant le plan Energie-Climat à atteindre d'ici à 2020 les objectifs des "3×20" qui visent :

- l'intégration de 20% d'énergies renouvelables dans le mix énergétique européen,
- la réduction de 20% des émissions de gaz à effet de serres (GES) par rapport à 1990,
- l'augmentation de 20% de l'efficacité énergétique par rapport à 1990.

Au plan national, cet engagement se traduit par la volonté de la France de réduire les émissions de GES de 1990 de 20% en 2020 et de 75% en 2050 (facteur 4), et de porter à 23% la part des énergies renouvelables dans le mix énergétique final d'ici 2020.

Au delà des engagements pris par les pays européens pour réduire leur consommation énergétique et promouvoir les énergies renouvelables, on peut distinguer quelques unes des préoccupations principales des clients finaux s'agissant des questions énergétiques. Ces préoccupations ont une intensité variable selon le type de consommateur, ses activités, ses usages de l'énergie. Parmi ces préoccupations se retrouvent le prix et la facture. Globalement, les utilisateurs attendent des prix les plus bas possibles et une facture compréhensible, présentant un détail des informations utiles. La maîtrise de l'énergie est l'un des principaux vecteurs pour réduire la facture énergétique. Elle passe par des investissements sur les bâtis, sur les processus de construction, sur les différents appareils, afin d'optimiser l'efficacité énergétique ainsi que par une gestion plus efficace de l'énergie en phase d'exploitation des bâtiments.

Aujourd'hui, la facture énergétique des bâtiments représente près de 44% de la facture globale en France, évaluée en 2012 à 69 milliards d'euros, soit 30,36 milliards d'euros (+11% par rapport à l'année précédente). L'efficacité énergétique des bâtiments représente donc un réel enjeu.

Les bâtiments sont les principaux leviers d'action d'optimisation de l'efficacité énergétique et de réduction des émissions de CO_2 dans les villes. Selon une étude sur les bâtiments, publiée en 2012 par l'ADEME (Agence De l'Environnement et de la Maîtrise de l'Energie), le secteur du bâtiment en France est le plus gros consommateur final d'énergie :

- le secteur résidentiel-tertiaire compte pour 44% du total de la consommation finale d'énergie, contre 32% pour les transports et 21% pour l'industrie, en 2011.

— le secteur résidentiel-tertiaire est le second émetteur de CO_2 en France avec 25% des émissions de CO_2.

La multitude des parties prenantes (promoteurs, constructeurs, propriétaires, occupants, exploitants ou gestionnaires) rend le secteur du bâtiment complexe. Il offre cependant de nombreuses potentialités de progrès et d'améliorations. Des efforts, émanant de la volonté politique à différentes échelles, ainsi que de nouvelles réglementations visant une meilleure efficacité énergétique des bâtiments ont émergé ces dernières années. Aujourd'hui, pratiquement tous les bâtiments neufs sont conçus en respectant les normes et labels de construction visant une meilleure efficacité énergétique : HQE (Haute qualité Environnementale), BBC (Bâtiment Basse Consommation), BEPOS (Bâtiment à Energie Positive), ...

En France, la nouvelle réglementation thermique en vigueur (RT 2012) prévoit de diviser par 3 la consommation d'énergie primaire des bâtiments neufs, labellisés « bâtiments basse consommation »

L'efficacité énergétique d'un bâtiment s'inscrit dans une approche globale qui s'articule autour de 4 piliers :

— L'efficacité énergétique passive (la Qualité du Bâti) : les matériaux et le type d'architecture sont pensés et adaptés à l'environnement et aux conditions extérieures. L'enveloppe du bâtiment est un moyen d'optimiser les échanges avec l'extérieur (réduction des pertes d'énergie par une meilleure isolation, renouvellement de l'air intérieur, utilisation de la lumière naturelle ...).
— L'efficacité énergétique active : il s'agit de la gestion intelligente (contrôle commande, pilotage et supervision) des consommations en fonction de la présence et des besoins des occupants. Celle-ci est assurée par la gestion des infrastructures techniques ainsi que par la régulation des équipements, par usage et par bâtiment (typiquement au travers d'un système de GTB (Gestion Technique du bâtiment)).
— Les équipements performants : ils visent une amélioration de la production et de la conversion d'énergie.
— L'implication des consommateurs.

En combinant l'amélioration de l'efficacité énergétique passive avec des technologies actives de gestion du bâtiment et de l'énergie, les bâtiments vont constituer une ressource additionnelle de flexibilité pour la gestion de l'énergie. L'amélioration de l'efficacité passive et active des bâtiments joue donc un rôle essentiel dans l'atteinte de ces objectifs.

Les systèmes passifs (architecture du bâtiment, sélection des matériaux adéquats) permettent d'améliorer le confort thermique des infrastructures. L'inertie thermique des bâtiments par exemple, permet de stocker l'énergie reçue par ces derniers et de la restituer progressivement quand cela est nécessaire. Moins sensible aux changements climatiques extérieurs, le bâtiment conserve une température relativement stable et permet de différer l'utilisation des systèmes de chauffage ou de climatisation.

L'efficacité énergétique active agit sur l'optimisation des flux énergétiques via l'utilisation d'équipements performants et de systèmes intelligents de comptages, de contrôle et de régulation. Ces derniers permettent la mesure des consommations, leur gestion et leur optimisation tout en prenant en compte les données internes et externes du bâtiment.

Un modèle de bâtiment tend à décrire les différentes caractéristiques de l'architecture, choix

des matériaux et le comportement des différents flux énergétiques opérant sur lui.

L'analyse de sensibilité peut aider à optimiser l'efficacité active et passive d'un bâtiment. Elle a été largement utilisée pour étudier les performances thermiques de bâtiments aussi bien du point de vue de la conception ([65],[62]), de la calibration de modèle ([122],[115]) que de l'impact du changement climatique ([118],[28], [50]). Cette analyse pourra avoir aussi comme but final d'aider l'utilisateur à réagir à un signal extérieur afin de modifier à la hausse ou à la baisse sa consommation énergétique. C'est-à-dire pouvoir avoir une certaine flexibilité énergétique. Elle peut être valorisée sur les marchés de l'électricité ou bien contribuer à lisser une courbe de consommation à une échelle locale, régionale ou nationale.

On peut remarquer que la plupart des études rencontrées dans la littérature ([38], [30],[112],[86]) se sont déroulées dans un cadre statique. Il est évident que l'efficacité active ne peut être étudiée à un instant donné. On peut suivre le même raisonnement pour l'efficacité passive. Le bâti, les matériaux ne réagissent pas de la même manière suivant les heures de la journée (ensoleillement, pluie, ...) ou les saisons par exemple.

Un second problème qu'il n'est pas évident de gérer lors d'une analyse de sensibilité est la dépendance des paramètres entre eux. Les méthodes d'analyse de sensibilité sont relativement bien définies lorsque les entrées du modèle sont indépendantes. Même s'il commence à se développer des méthodes dans un cadre d'entrées dépendantes le sujet reste relativement ouvert ([17], [76]). Nous avons proposé dans les chapitres précédents des solutions adaptées aux méta-modèles dynamiques que nous allons appliquer à l'étude d'un bâtiment. Ce sujet est resté auparavant très peu traité dans un cadre dynamique.

Le dernier problème qui n'est pas des moindres est la modélisation des entrées. Les propriétés thermo-physiques de l'enveloppe du bâtiment sont le plus souvent modélisées par des distributions normales ou uniformes sans étude qualitative suffisante. Les flux énergétiques quant à eux peuvent être modélisés par des séries temporelles.

Chapitre 2

Modélisation d'un bâtiment

L'approche la plus classique pour modéliser un bâtiment est de se baser sur des considérations physiques. La formulation du modèle est basée sur la connaissance des caractéristiques physiques des matériaux utilisés et de sous modèles représentant la conduction thermique, la convection et le rayonnement.

C'est le cas des modèles obtenus par logiciel COMFIE, TRNSYS, EnergyPlus [97].

COMFIE : Ce logiciel repose sur une approche physique et permet d'évaluer avec précision l'influence de l'inertie et des apports thermiques sur les consommations. Il s'appuie sur un modèle aux différences finies multi-zone réduit par analyse modale.

COMFIE décompose le bâtiment en mailles, supposées à température uniforme :

- des mailles représentant des portions de parois. Pour cela, la composition de chaque paroi est analysée afin de déterminer la disposition et le nombre de mailles requis.
- une maille supplémentaire caractérisant l'ambiance intérieure de chaque zone thermique.

La finesse du maillage est un compromis entre le temps de calcul et la précision du résultat désirée.

Pour la mise en équation, un bilan thermique est déterminé sur chacune des mailles afin d'évaluer tous les flux de chaleur échangés.

Les caractéristiques (conductivité, capacité thermique et masse volumique) de tous les matériaux constitutifs des parois sont prises en compte, ce qui est beaucoup plus détaillé qu'une description de paroi par un coefficient d'échange thermique et une classe d'inertie (légère, moyenne, lourde).

Les modèles de chaque zone thermique sont couplés, afin de constituer un modèle global du bâtiment. Il est pour cela nécessaire de décrire les parois internes séparant les différentes zones. Chaque modèle de zone est réduit par analyse modale. Cette simplification permet de ne calculer que les modes les plus représentatifs de l'évolution dynamique des composants (murs, planchers, ...) ou des sollicitations (variations d'ensoleillement, puissance de chauffage régulée, ...).

La finesse du résultat dépend du temps de calcul, et donc du nombre de constantes de temps conservées.

Un grave inconvénient de l'approche traditionnelle est la difficulté d'obtenir un paramétrage raisonnable. En général, le modèle global final possède un assez grand nombre de paramètres, et, en raison des approximations et des simplifications inévitables, présentes à la fois dans les modèles de chacun des sous-processus individuels et dans le couplage entre les différents sous-processus, il est très difficile de prédire la précision du modèle global.

Circuit électrique : On a l'habitude de représenter de manière analogique par un signal électrique les variations de grandeurs physiques diverses. C'est pourquoi une seconde manière de modéliser un bâtiment est un circuit électrique [46] (figure : III.1).

Les températures ambiantes des différentes pièces sont associées à des générateurs de tensions (des sollicitions). Les différents apports thermiques tels que le chauffage, le flux solaire, les personnes à l'intérieur sont représentés par des générateurs de courant. Les caractéristiques physiques du bâtiment sont définies par des composants électriques : résistances, capacités et sont à déterminer. Toutes les températures sont mesurées à l'exception des températures internes des murs qui sont à estimer.

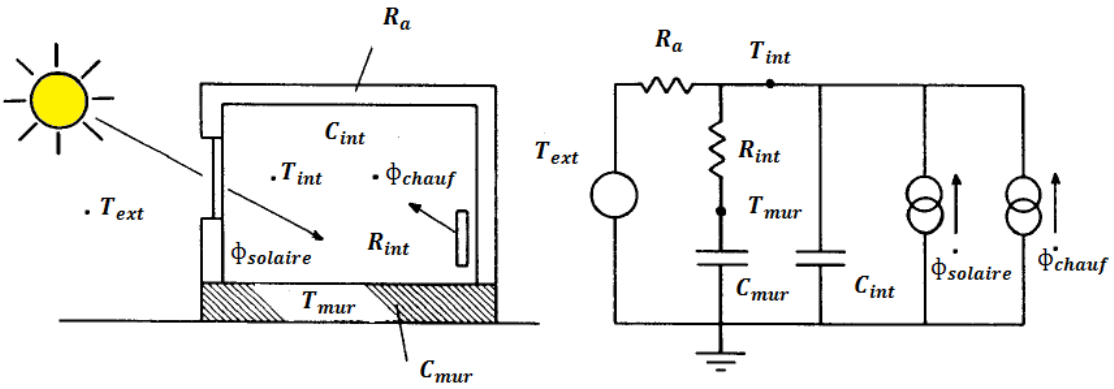

FIGURE III.1 – Modélisation par circuit électrique

Cette représentation peut aussi se ramener à une représentation d'état. Les températures internes de murs correspondent alors aux variables d'état. Cette représentation joue un rôle important en identification et d'une manière générale dans la théorie de l'estimation des paramètres de modèles et des variables d'état inconnues ([64],[120], [51]). Le problème du bruit évoqué plus loin et les difficultés des méthodes d'estimation semblent être des causes de limitation en ce qui concerne au moins les études de sensibilité.

Méthode liées à l'optimisation : Le dernier modèle déterministe que l'on peut considérer est un modèle d'état où l'on estime les paramètres des matrices sans considération physique. La plupart du temps ces paramètres sont estimés par des méthodes d'optimisation déterministe

classique. Le plus simple est une analyse de régression non linéaire par rapport aux données d'observation et les résultats du modèle de calcul. Dans ce cas, le problème est résolu en utilisant des techniques d'optimisation afin de minimiser la somme des carrés des résidus entre les prédictions du modèle et les données observées ou expérimentales.

Cette approche peut être utile pour une estimation déterministe initiale des variables et des paramètres et peut conduire à une première estimation (ajustement du modèle direct) avec un nombre limité d'essais de la fonction objective de régression non linéaire. Cette formulation intuitive, présente plusieurs inconvénients :

— le temps de calcul de ces algorithmes est long surtout si le nombre de paramètres est important
— il n'y pas forcément unicité de la solution. Plusieurs valeurs peuvent conduire à la même minimisation.

De plus si l'on utilise un algorithme d'optimisation déterministe le résultat dépend souvent des valeurs initiales et donc on n'est pas assuré de trouver le minimum global.

Le but de ces modèles étant la prédiction, il n'est pas forcément intéressant d'avoir les meilleures valeurs des paramètres permettant de minimiser l'erreur avec les données utilisées. On préfère donc rentrer des valeurs initiales avec un a priori physique et obtenir un minimum à partir de ces conditions ([51], [120]). On peut sinon pour palier à ce défaut utiliser des méthodes de ré-échantillonnage (Monte Carlo, Hyper Cube Latin). Le temps de calcul sera néanmoins aussi très important car nécessitant de nombreux runs.

Modèle stochastique : On peut distinguer deux sortes de travaux de modélisation stochastique concernant des bâtiments :

— des travaux visant à la prédiction et utilisant des méthodes de réduction de modèles si les paramètres et/ou les variables d'entrée sont trop nombreux.
— des travaux étudiant la sensibilité et/ou la propagation des incertitudes.

Dans ces travaux, essentiellement ceux du premier groupe, la dimension temporelle est prise en compte par l'utilisation de modèles $VARMAX$ et des modèles d'état.
La dynamique dans tous les cas est engendrée pour l'entrée par un bruit et la relation entrée-sortie est donc linéaire.
Contrairement à certains modèles physiques les modèles d'état sont bruités notamment l'équation d'observation. Introduire des bruits dans un modèle semble justifié pour les raisons presque évidentes suivantes :

— les matrices d'état choisies ne sont qu'une approximation et donc ne représentent pas le "vrai" comportement du système thermique
— certaines variables ou entrées ne sont pas connues ou non pas été prises en compte (par exemple le vent ou certains flux solaires)
— les données utilisées sont des données mesurées. Elles ne sont, par définition, pas exactes (même si considérées comme telles), d'où la notion de mesures bruitées.

Les travaux les plus nombreux sont ceux du premier groupe, donc hors étude de la sensibilité. Tous utilisent des modélisations $VARMAX$, le plus souvent $VARX$ et les comparent souvent à la variante représentation d'état. De même les méthodes classiques d'estimation par moindres carrés ou vraisemblance sont comparées aux méthodes qui utilisent le filtre de Kalman souvent difficile à mettre en œuvre. La plupart des travaux utilisent la discrétisation d'équation

différentielles représentant des échanges de chaleur. Citons en premier Madsen et al. [81] qui part de l'équation différentielle liant une température intérieure avec la température extérieure, l'entrée chauffage et la radiation solaire mesurée sur un mur vertical orienté au sud. La comparaison est faite aussi avec le modèle électrique classique.

J. Pakanen et al. [90] ont choisi des modèles $VARMAX$ pour représenter un modèle de chauffage par exemple. C'est aussi le cas de L. Ferkl et al. [44] qui ont comparé les avantages de la modélisation de la température d'un bâtiment test par modèle d'état ou par un modèle $VARMAX$. Le modèle d'état s'est révélé plus simple à implémenter, rapide à estimer et précis lorsque le bruit du système est blanc. Cependant les modèles $VARMAX$ obtiennent de meilleurs résultats pour les systèmes dans lesquels les paramètres d'entrées ont des bruits respectifs loin d'être idéaux. Ils semblent donc mieux adaptés à la réalité des études sur le bâtiment.

Chapitre 3

Analyse de sensibilité et bâtiment

L'analyse de sensibilité vise à aider à optimiser l'efficacité active et passive d'un bâtiment. Elle a été largement utilisée pour étudier les performances thermiques de bâtiments aussi bien du point de vue de la conception ([65],[29]), de la calibration de modèles ([92],[115]) que de l'impact du changement climatique ([28]).

Afin d'optimiser les performances énergétiques pour ces différentes études il est utile d'évaluer l'impact des sources d'incertitudes qui peuvent influencer ces performances. Ainsi se distinguent deux problèmes importants :

1. Comment modéliser ces sources et rendre compte de leur caractère incertain ?
2. Comment quantifier ou mesurer leur impact sur la variable d'intérêt ?

Nous avons vu dans la partie précédente les différents modèles à notre disposition. Aussi il est important de rappeler que le choix d'un modèle est déterminé par ce que l'on souhaite montrer ou représenter et les hypothèses entourant la modélisation sont des choix sous réserve de justifications qui nous sont propres.

Le premier choix est la variable d'intérêt (sortie). Dans notre cas, ce sera la température intérieure d'une pièce qui représente un critère de confort. Cela pourrait être aussi la consommation totale d'énergie par exemple. Concernant les paramètres et les variables qui influencent cette sortie, les questions que l'on peut se poser sont : de quoi sommes nous sûrs ? Quelles connaissances avons nous sur ces variables ou paramètres ? Comment modéliser cette incertitude ?

Par exemple, dans Eisenhower B. et al. [38], les auteurs ont choisi d'étudier dix paramètres relatifs à la consommation énergétique et électrique du bâtiment en fonction des paramètres relatifs à la géométrie du bâtiment, le système d'eau, de chauffage et d'électricité conçu sur EnergyPlus [1].

On peut connaitre un point de fonctionnement (par exemple une température de consigne de chauffage), grâce à des a priori physiques ou des mesures. On va alors faire une étude dite locale, c'est-à-dire étudier l'impact de petites variations sur la sortie. L'analyse de sensibilité locale a été aussi largement utilisée dans le domaine de l'analyse de la consommation d'énergie d'un bâtiment ([77], [45], [80]).

1. http ://apps1.eere.energy.gov/buildings/energyplus/

Par exemple Demanuele et al. [30] ont utilisé une analyse de sensibilité différentielle pour déterminer les facteurs clés qui influent sur la consommation totale d'énergie dans une école au Royaume-Uni. Les variables importantes sont liées aux occupants, telles que la charge de l'équipement (bureau, matériel scolaire), les heures de présence, l'horaire de chauffage et les températures de consigne. C'est le cas aussi de Lam et al. [77] qui ont étudié la performance énergétique d'un immeuble de bureaux à Hong Kong. Les résultats indiquent que la consommation d'énergie annuelle est très sensible aux charges internes, à la température des points de réglages, et l'efficacité de l'équipement de ventilation.

Les modèles linéaires étant largement répandus dans la modélisation de bâtiments, l'indice de sensibilité SRC a été largement utilisé [32]. Par exemple Hygh et al. [65] ont exploré la performance énergétique d'immeubles de bureaux dans quatre villes des Etats-Unis en utilisant l'indice de sensibilité SRC. Les résultats montrent que l'influence des paramètres de conception sur l'utilisation de l'énergie dans un bâtiment varient selon les différentes zones climatiques. Ballarini et al. [6] ont utilisé l'indice SRC pour déterminer les variables-clés qui influent sur l'énergie de refroidissement d'un immeuble résidentiel en Italie. Les résultats montrent que les trois premiers facteurs les plus importants sont la protection solaire, la surface des fenêtres, et l'isolation des fenêtres. Breesch et al. [11] appliquent cet indice pour déterminer les facteurs les plus influents affectant le confort thermique d'un immeuble typique de bureaux en Belgique. Ils ont constaté que les gains de chaleur internes et l'étanchéité à l'air sont les deux variables les plus importantes.

Le deuxième indice adapté dans le cadre non linéaire est l'indice de Sobol. Spitz et al. [112] ont utilisé une méthode d'estimation Pick and Freeze afin d'identifier les paramètres les plus influents d'une maison expérimentale en France. Les six facteurs importants qui affectent la température de l'air ont été proposés : la capacité de chauffage, l'infiltration, l'épaisseur de la laine de verre, l'efficacité de l'échangeur de chaleur, les gains de chaleur internes et la conductivité de la laine de verre. Mechri et al. [86] ont mis en œuvre une méthode FAST pour identifier les variables clefs affectant la performance thermique d'un bâtiment typique de bureaux en Italie. Les résultats indiquent que l'enveloppe est le facteur le plus important à la fois pour le chauffage et l'énergie de refroidissement.

Tous les paramètres étudiés dans la plupart de ces travaux sont des paramètres statiques (ils n'évoluent pas au cours du temps ou du moins cela est supposé par les auteurs). Il y a cependant des variables dynamiques, telles que la température extérieure qui peuvent influencer la variable d'intérêt. Une solution envisagée par [90] est de modéliser ces variables par séries temporelles. Ces entrées considérées comme des processus aléatoires utiliseront une analyse de sensibilité globale.

Une remarque importante à considérer est que peu de ces études prennent en compte le fait que tous ces paramètres ou variables sont dépendants entre eux.

Chapitre 4

Positionnement du problème, données et modélisations

Depuis 2008 les pays européens se sont engagés à réduire leur consommation énergétique dans le cadre du plan Energie-Climat. L'un des premiers leviers d'action afin d'optimiser l'efficacité énergétique est le secteur du bâtiment. L'efficacité énergétique est obtenue à partir d'une vision globale pouvant se résumer en quatre axes principaux : l'utilisation d'équipements performants, une implication des consommateurs, la gestion d'une efficacité dite passive (qualité du bâti) et une efficacité active autrement dite intelligente gérée en fonction des besoins.
Nous cherchons à illustrer ici que l'utilisation des méthodes d'analyse de sensibilité peuvent participer à optimiser l'efficacité active ou passive d'un bâtiment. Dans ce travail nous utiliserons la méthode d'analyse de sensibilité temporelle, pour deux modèles décrivant l'impact des changements, notamment les températures des pièces environnantes et la température extérieure, sur le confort d'une pièce (sa température intérieure) et sur la puissance de chauffage fournie. On cherche donc à savoir quelles sollicitations des pièces voisines (températures, apports thermiques ou paramètres : isolant, épaisseur des murs, inertie thermique, surface de vitrage) influencent le plus le confort des usagers ou la puissance fournie pour maintenir ce confort.

Proposer un diagnostic quant au problème de gestion énergétique nécessite un modèle. Les modèles statistiques et les modèles mécanistes qui consistent ici à expliquer les processus dynamiques des échanges thermiques sont développés à partir d'observations. Pour cela, nous avons à notre disposition une plateforme expérimentale : Predis MHI [1], comportant une salle de classe accueillant des étudiants durant l'année scolaire. La température de cette salle est donc sensible aux jours de la semaine, aux heures mais encore aux périodes de vacances, saisons et météo. La variable d'intérêt, la sortie de notre modèle, est la température intérieure de cette salle ou la puissance de chauffage. Elles sont notées T^{int} et K. Le vecteur $(U_t)_{t\in\mathbb{N}}$ contient les sollicitations, définies aux différents instants t exprimés en heure. Ce sont les variables d'entrées dynamiques. Le vecteur θ contient les paramètres du bâtiment qui représentent les caractéristiques physiques des matériaux utilisés.

Dans la suite, nous allons présenter les variables utilisées, notamment les différentes tempé-

1. http ://predis.grenoble-inp.fr/smartbuilding

ratures, le chauffage et l'apport thermique des usagers. Nous détaillerons les traitements que l'on a utilisés afin de créer des modèles entrée-sortie et des modèles de simulation des entrées. En dernière partie nous présentons les résultats d'analyse de sensibilité appliquée à ces différents modèles (entrées-sorties). Nous avons choisi dans un premier temps de ne pas étudier en détail la sensibilité par rapport aux paramètres physiques et de nous concentrer sur l'aspect dynamique des modèles. Pour rappel, la sensibilité des paramètres relève plutôt d'une étude de type statique.

4.1 Les variables

4.1.1 Températures

La pièce où évolue la variable d'intérêt T^{int} est entourée d'un couloir, d'un bureau adjacent, d'un bureau qui se trouve au dessous et d'un puits de lumière (un shed) (figure : III.1). Ces différentes pièces sont équipées de capteurs de température depuis 2009. Les températures, mesurées toutes les heures, seront notées :

$$U_t = (T^{\text{cor}}, T^{\text{off}}, T^{\text{above}}, T^{\text{below}}, T^{\text{ext}})$$

La température extérieure notée T^{ext} est également mesurée.

En ce qui concerne les températures, la qualité des données est assez médiocre. Il existe de nombreuses plages horaires creuses dues à de nombreux arrêts du système d'acquisition. Par exemple, l'année 2011 est inexistante (voir figure III.2). En 2011, un nouveau serveur dédié aux mesures a été installé non sans difficulté. Les problèmes ont persisté malgré cette modification. La surveillance constante du système est nécessaire pour un redémarrage manuel en cas de panne ce qui explique les trous de plusieurs heures voire plusieurs jours. La plupart des plages horaires manquantes sont supérieures à 8 heures il est donc difficile de combler les vides en utilisant des logiciels traitant des données manquantes en utilisant par exemple l'algorithme EM [85].

Aujourd'hui un système d'alerte automatique a été mis en place avec une relance automatique de tout le système ce qui nous laisse supposer que les prochaines données seront plus exploitables.

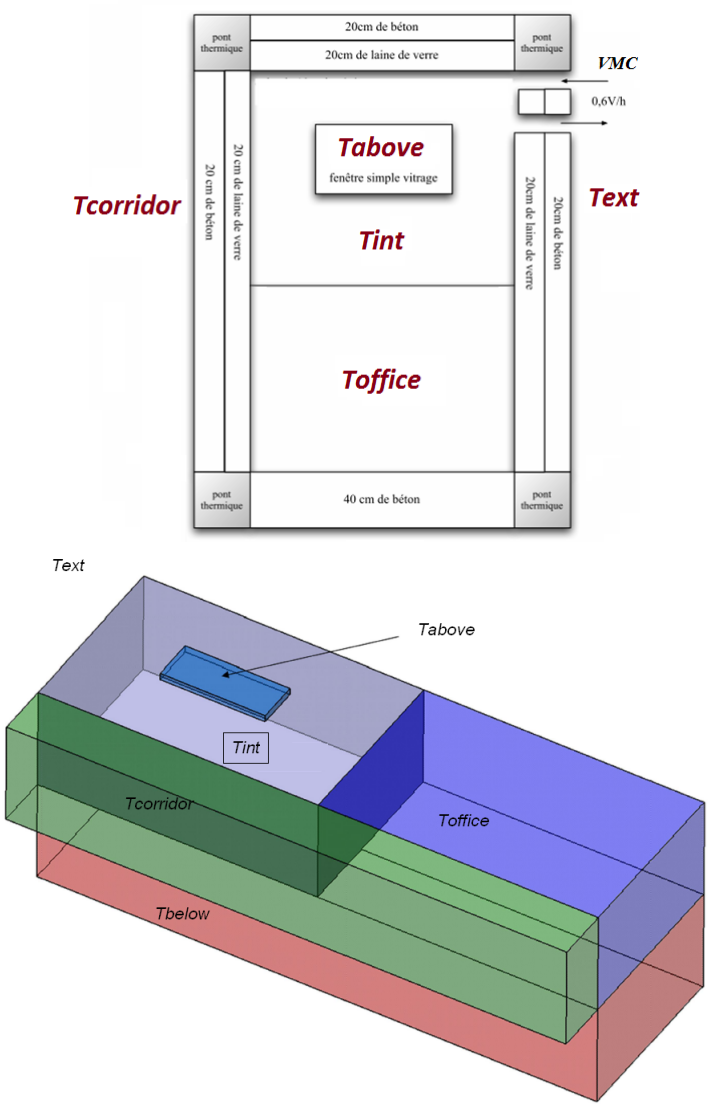

FIGURE III.1 – Agencement des différentes pièces

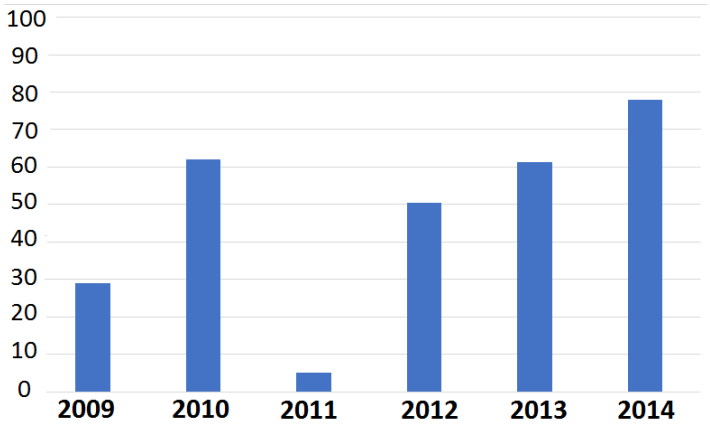

FIGURE III.2 – Pourcentage de données disponibles par année concernant les températures

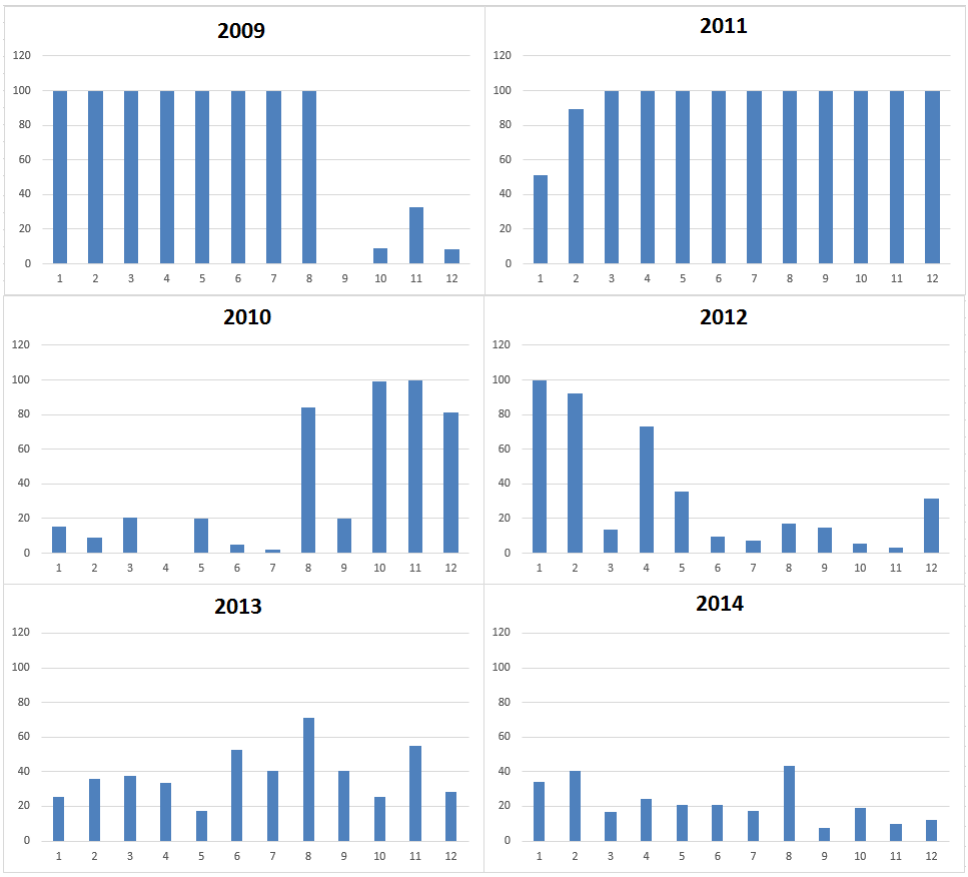

FIGURE III.3 – Pourcentage de données de températures manquantes par mois

Nous avons alors décidé de travailler mois par mois. Nous ne possédons que très peu de mois complets sans données manquantes (figure III.2, III.3 et III.4). De plus le fait que le chauffage ne fonctionne que 6 mois par an, conduit à faire deux modèles : un modèle été et un modèle hiver. Les mois exploitables ne sont pas nécessairement ceux de l'année suivante. Ceci implique des difficultés sur l'estimation des saisonnalités.

Les mois choisis pour l'étude sont fonction du protocole de chauffage qui a varié au cours du temps. La plage étudiée pour le modèle été est du 22/08/2012 au 04/09/2012 qui correspond à la plage la plus grande ne souffrant pas d'un manque grave de données en ce qui concerne les températures.

La plage étudiée concernant le modèle hiver est celle du 13/11/2012 au 13/12/2012. Cette année ne comporte que deux mois où le chauffage a été mesuré ce qui réduit considérablement les plages d'étude (figure III.6). Cette période a été choisie parce que les données de températures sont complètes.

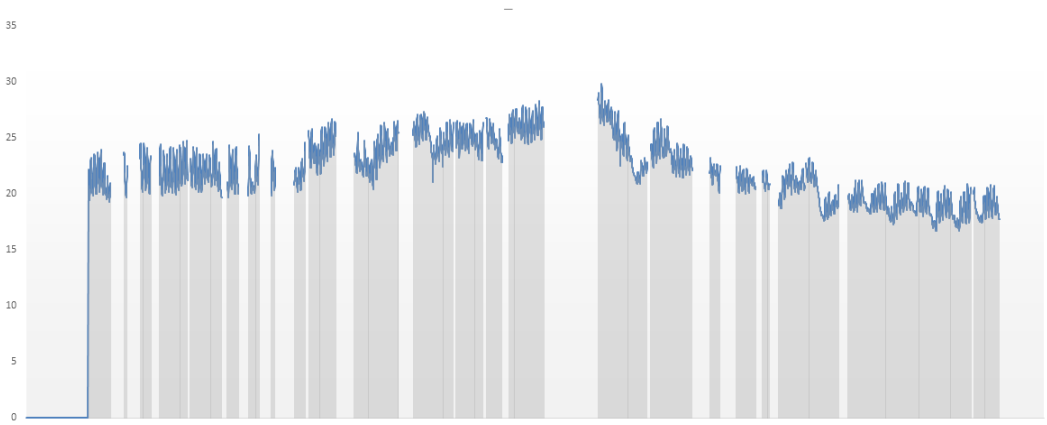

FIGURE III.4 – Température du couloir en 2012

4.1.2 Chauffage

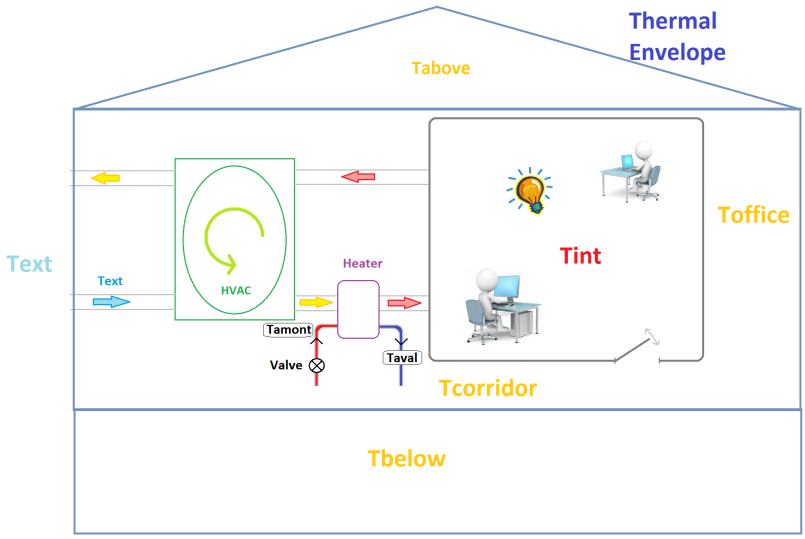

FIGURE III.5 – Représentation schématique du bâtiment et de la VMC double flux

La pièce où évolue T^{int} est chauffée grâce à une VMC (ventilation mécanique contrôlée) double flux (figure : III.5). Ce dispositif mécanique permet de récupérer la chaleur de l'air prélevé dans la salle et de la transmettre via un échangeur à l'air prélevé de l'extérieur. Cet air ainsi préchauffé est ensuite chauffé grâce à une batterie d'eau chaude. Le chauffage sera noté K_t. Il est mesuré en Watt toutes les heures et n'est présent que lors des mois d'Octobre à Mars.
Le chauffage est déterminé à partir d'un bilan thermique de la batterie d'eau chaude :

$$K_t = D_e.\%Vanne.C_e.(T_{amont} - T_{aval}) \tag{III.1}$$

— T_{amont} : la température d'eau en amont de l'échangeur eau-air et T_{aval} la température d'eau en aval de l'échangeur eau-air.
— D_e : le débit nominal (ouverture de vanne 100%) d'eau : $D_e = 0.0082kg/s$.
— $\%Vanne$: l'ouverture de la vanne d'eau de la batterie d'eau chaude.
— C_e : la chaleur massique de l'eau $C_e = 4000J.kg^{-1}.°C^{-1}$.

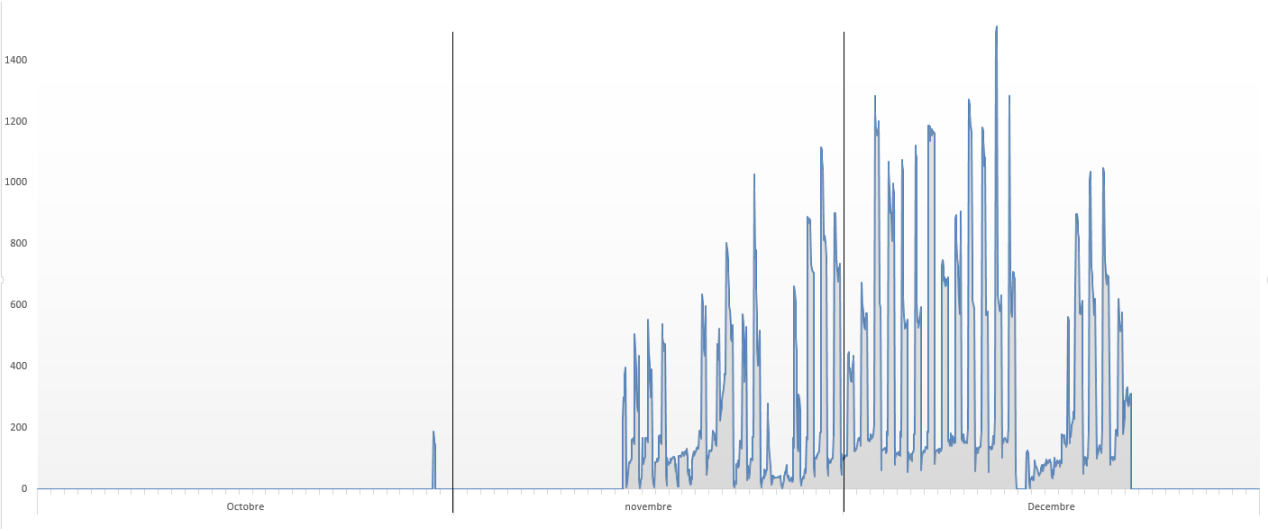

FIGURE III.6 – Mois de fonctionnement du chauffage en 2012

Le protocole de débit d'air de la VMC conditionne le flux de chauffage.
Ce débit n'a pas été maintenu constant plus d'un mois. Il ne possède aucune régularité.

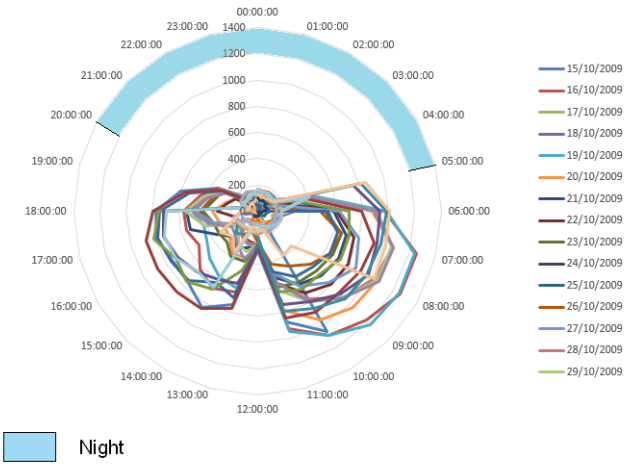

FIGURE III.7 – Représentation sur une journée du chauffage en Watt du 15/10/2009 au 15/11/2009

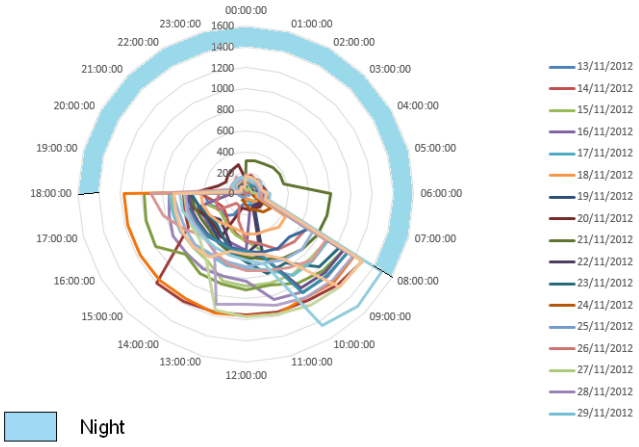

FIGURE III.8 – Représentation sur une journée du chauffage en Watt du 13/11/2012 au 13/12/2012

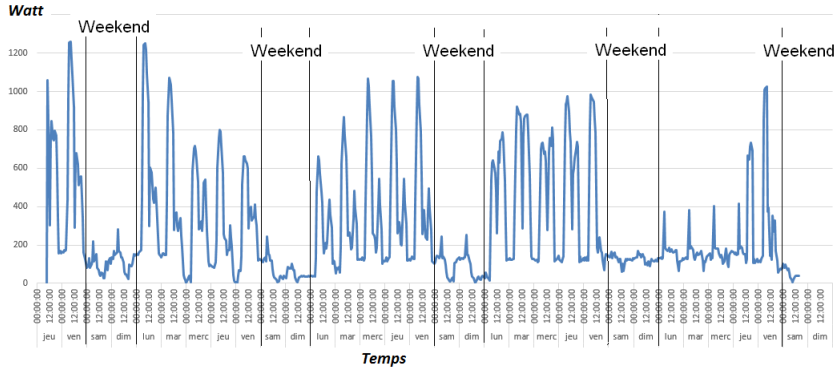

FIGURE III.9 – Flux thermique en Watt du 15/10/2009 au 15/11/2009

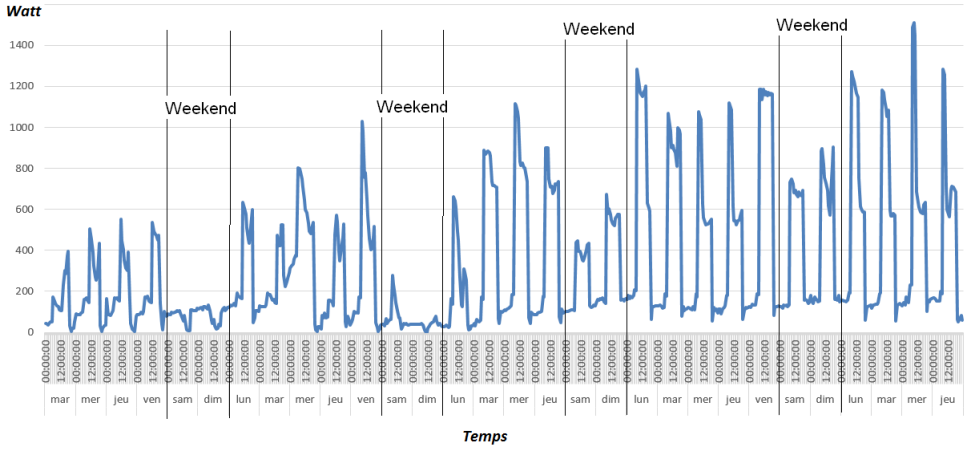

FIGURE III.10 – Flux thermique en Watt du 13/11/2012 au 13/12/2012

Le protocole a changé en 2012. Le chauffage fonctionnait de 6 heures à 19 heures et était coupé à midi (figure : III.7), il ne fonctionnait pas le week-end. En 2012-2013 le chauffage fonctionnait le midi et parfois les week-ends (figure : III.8).

Pendant les mois disponibles pour notre étude, le débit fixé par l'utilisateur n'a pas été maintenu constant (figure : III.9). Les week-ends, le débit, a été parfois réactivé. Ces modifications sont visibles sur la figure : III.10 du flux de chauffage, celui-ci étant proportionnel au débit.

4.1.3 Présence des personnes : équivalent chaleur

La présence des étudiants est calculée en équivalent chaleur. Il est obtenu grâce à un capteur de CO2, installé en 2012, qui recueille des mesures toutes les minutes. Cette équation exprime le taux d'accroissement de CO_2 de la pièce. Le CO_2 est produit par les occupants et dépend du renouvellement en air neuf effectué par la VMC.

$$N_t = \frac{((CO_2)_t - (CO_2)_{t-1})V/\Delta t + D_{ventilation}((CO_2)_t - (CO_2)_{air})}{D_{personne}(CO_2)_{personne}} \tag{III.2}$$

- $(CO_2)_t$: concentration de CO_2 à l'instant t en ppm
- V : volume de la pièce en m^3
- $D_{ventilation}$: débit d'air soufflé dans la pièce en $m^3.h^{-1}$
- $(CO_2)_{air}$: concentration de CO_2 dans l'air en ppm
- $D_{personne}$: débit de CO_2 expulsé par une personne en $m^3.h^{-1}$
- $(CO_2)_{personne}$: concentration de CO_2 dans l'air expiré par une personne en ppm
- Δt : durée exprimée en heure.

Afin de palier à certaines données aberrantes, les données sont lissées sur une heure. En effet si quelqu'un respire à côté du capteur cela fausse l'estimation : une grosse quantité de CO2, qui n'a pas encore eu le temps de se mélanger à l'air de la pièce est mesurée.

La présence des étudiants est observée les jours ouvrés de 8h à 18h. Il n'y a personne la nuit entre 18h et 8h du matin ainsi que les week-ends.

Les capteurs ayant été installés en 2012, nous ne possédons qu'un seul mois où sont représentées toutes les entrées (figure : III.11).

Nous étudierons alors le mois de novembre du 13/11/2012 au 12/12/2012.

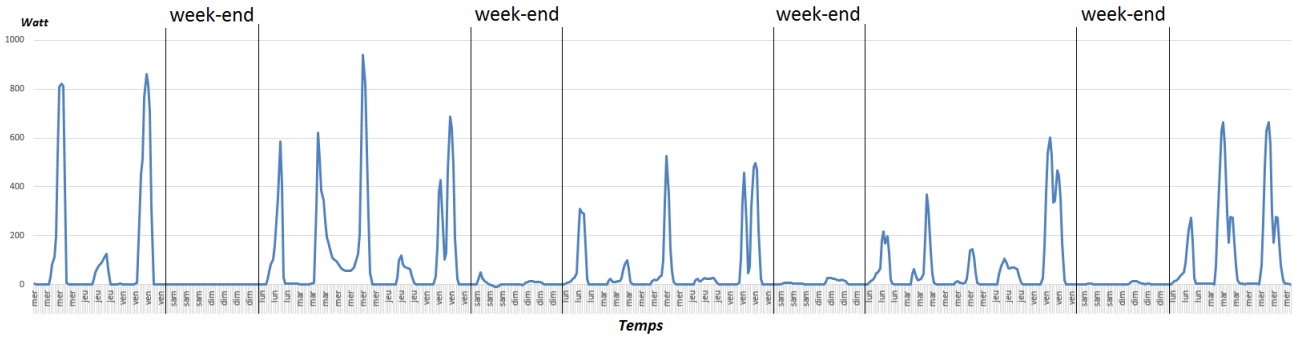

FIGURE III.11 – Equivalent chaleur du nombre de personnes en Watt en fonction du temps du 13/11/2012 au 12/12/2012

L'ensemble des sollicitations prises en compte dans les différents modèles utilisés pour décrire la variable d'intérêt sont résumées dans le tableau 4.1.

Notations	Définition
T^{int}	Température intérieure de la pièce centrale
T^{ext}	Température extérieure
T^{cor}	Température du couloir
T^{off}	Température du bureau adjacent à la pièce centrale
T^{above}	Température de la pièce se situant en dessus de la pièce centrale
T^{below}	Température de la pièce se situant en dessous de la pièce centrale et du bureau
K	Puissance de chauffage utilisée pour chauffer la pièce centrale
N	Equivalent chaleur du nombre de personnes présentes dans la pièce centrale

TABLE 4.1 – Définition des variables de la relation environnement-confort et environnement-chauffage

4.2 Modèles entrée-sortie

Les modèles entrée-sortie que l'on a développés sont des modèles matriciels à temps discret (le pas de temps est horaire). Le premier modèle dynamique utilisé dans ce travail vise à décrire l'évolution de la température intérieure de la pièce centrale T^{int} en été en fonction des différentes températures. Le second concernant la période hiver étudie le comportement de la puissance du chauffage en fonction des différentes températures ainsi que l'équivalent chaleur des visiteurs. Toutes les entrées de ces modèles sont dépendantes entre elles. Nous avons vu aux chapitres 4 et 2 de la première partie que dans le cas de variables dépendantes la situation est complexe ; en effet toutes les variables liées à X^1 vont intervenir pour calculer l'indice de sensibilité S^{X^1}. Lorsque les variables sont très fortement corrélées comme c'est le cas des variables intervenant dans l'étude en termes de sensibilité de notre bâtiment test,

les confusions d'effet rendent les interprétations difficiles et posent le problème du choix du système d'entrée-sortie. Doit-on se limiter à un seul choix de variables d'entrées ?

Prenons l'exemple suivant. On dispose comme données de quatre températures de salle, de la température extérieure, du chauffage, de l'équivalent chaleur et comme données de sortie la température intérieure d'une salle privilégiée T^{int}. On peut proposer des modèles du type :

- $Y_t = T^{\text{int}} = \eta_1(4 \text{ températures de salle, chauffage, équivalent chaleur})$
- $Y_t = T^{\text{int}} = \eta_2(4 \text{ températures de salle})$
- $Y_t = T^{\text{int}} = \eta_3(T^{\text{ext}}, \text{ chauffage, equivalent chaleur })$

η_1, η_2, η_3 peuvent être construits de différentes manières, par exemple de manière stationnaires ou cyclo-stationnaires.

Si l'on prend le cas de η_2, seules les températures des salles intérieures adjacentes interviennent. Ce modèle permet sans confusion d'effets d'analyser les problèmes de conduction thermique à l'intérieur (seul) du bâtiment. Dès que le modèle fait intervenir en entrée la température extérieure, les qualités d'isolation des murs extérieurs seront étudiées. On se rend compte que le choix de modèle et la réduction de modèle ne sont pas les mêmes sujets. La réduction de modèle est souvent pensée en terme d'entrées indépendantes et l'on veut se débarrasser des entrées peu influentes. Ici dans le choix de modèle, au vu des fortes dépendances des facteurs candidats à être des entrées, il s'agit de comprendre quelles sont les décompositions de la variance de Y qui sont les plus utiles, c'est-à-dire les plus facilement interprétables, cela ne ressort pas d'un choix purement mathématique.

Les seuls outils dont on dispose du point de vue théorique semblent être les décompositions de Hoeffding dans le cas d'entrées dépendantes développées dans la partie Sensibilité de l'article de Chastaing et al. [17]. L'application numérique est très difficile et l'interprétation de l'interaction reste à conforter. Ces outils sont difficiles à utiliser ici à cause de la dimension temporelle.

Pour cet ensemble de raisons nous avons étudié plusieurs modèles entrée-sortie construits à partir du même jeu de données mais utilisant de manière différente l'information. La méthode d'estimation des indices de sensibilité que nous proposons ne repose pas sur la forme du modèle entrée-sortie, mais sur la forme du modèle d'entrée. Nous aurions pu développer des modèles entrée-sortie non linéaires plus complexes mais par simplicité nous avons choisi deux types de modèles linéaires pour Y (la sortie) : des modèles de types VAR et des régressions. Les modèles diffèrent uniquement en ce que le VAR est une régression infinie de tout le passé des entrées et dont les coefficients décroissent. La régression ne fait intervenir qu'un passé fini des entrées. Nous avons étudié à titre d'exploration des modèles de régression non linéaires mais le manque de données n'a pas permis une estimation fiable.

La relation entrée-sortie est ici étudiée comme un méta-modèle statistique dont la qualité première sera de permettre une simulation assez facile. Si

$$Y_t = \eta(\mathbb{U}_{t,k}, \theta)$$

est le modèle d'entrée-sortie avec $\mathbb{U}_{t,k} = (U_t, \ldots, U_{t-k}, 0 \leq k \leq \infty)$ et $\theta \in \mathbb{R}^k$. L'estimation de θ se fait à partir de paires temporelles $(\mathbb{U}_{t,k}, Y_t)_{t\in(0,T)}$, éventuellement avec plusieurs trajectoires d'apprentissage $(\mathbb{U}^{(i)}_{t,k}, Y^{(i)}_t)_{t\in(0,T),i\in\mathbb{N}}$.

Remarque 10. *Si $Y_t = \alpha Y_{t-1} + A(\theta)\mathbb{U}_{t,k}$, par exemple, c'est-à-dire que Y_t est donné avec une*

dynamique, le modèle se ramène au cas :

$$Y_t = (1 - \alpha d)^{-1} A(\theta) \mathbb{U}_{t,k} \tag{III.3}$$

Dans tous les cas, on peut utiliser la méthode des moindres carrés pour estimer les paramètres de III.3 et éventuellement α si le modèle est auto-régressif.

Remarque 11. *Supposons que le modèle d'observation soit bruité du type :*

$$Y_t = \eta(\mathbb{U}_{t,k}, \theta) + \varepsilon_t$$

avec ε_t indépendant de $\mathbb{U}_{t,k}$.
ε n'interviendra pas dans le calcul de la sensibilité (indices de Sobol). Dans ce cas même si ε_t n'est pas gaussien, on peut utiliser la vraisemblance Gaussienne comme contraste donnant un estimateur $\hat{\theta}_t$, et qui converge à partir de la minimisation de :

$$\frac{-1}{2} \sum_{t=0}^{T} \frac{|Y_t - \eta(\mathbb{U}_{t,k}, \theta)|^2}{\sigma^2} - \frac{T}{2} log(\sigma^2)$$

Les modèles entrée-sorties développés sont calculés à partir des données bruts.

4.2.1 Modèle été avec pour sortie la température intérieure T^{int}. Description générale

Le premier modèle été choisi est un modèle dynamique de type *VARX* que l'on notera **S1-A** :

$$T_t^{\text{int}} = \sum_{k=1}^{p} a_k T_{t-k}^{\text{int}} + \sum_{k=1}^{q} B_k \boldsymbol{U}_{t-k} \tag{III.4}$$

et le second un modèle de régression **S1-B** :

$$T_t^{\text{int}} = \sum_{k=1}^{q} B_k \boldsymbol{U}_{t-k} \tag{III.5}$$

Le choix des ordres p, q se fait de façon hiérarchique en fixant deux des ordres et en appliquant le critère de AKAIKE pour déterminer l'ordre par des essais successifs en étudiant les résidus. Dans le premier modèle (III.4) T^{int} dépend de tout le passé des entrées. La longueur du passé pris en compte dépend des paramètres α_k. A l'inverse dans le modèle de régression T^{int} dépend seulement des entrées $\boldsymbol{U}_{t-k}$ avec $1 \leq k \leq q$.

Il n'est pas rare que plusieurs hypothèses de modélisation a priori réalistes coexistent. Il est nécessaire alors de réaliser une analyse de sensibilité du modèle pour chaque application ce qui nous permet d'avoir un certain recul vis-à-vis du modèle créé.

Dans un second temps la forte corrélation 0.9 entre la température intérieure et T^{off} (table : A.0.1) peut entrainer des colinéarités qui peuvent fausser l'analyse de sensibilité. Cette pièce directement adjacente, possède une température pratiquement identique à celle de T^{int}. Nous

avons alors décidé de créer un modèle ne comportant pas T^{off}. On peut d'un point de vue physique considérer que la pièce centrale et le bureau n'en forme plus qu'une. Nous avons alors considéré un modèle de type *VAR* (**S2-A**) et de régression (**S2-B**) construit sur les entrées $(T^{\text{cor}}, T^{\text{above}}, T^{\text{below}}, T^{\text{ext}})$.

La mémoire des modèles retenus par critère AIC est de 2. Les coefficients des différents modèles sont résumés dans le tableau : 4.2.

	VAR **S1-A**	Régression **S1-B**	*VAR* **S2-A**	Régression **S2-B**
T^{int}_{t-1}	1.06	-	1.03	-
T^{int}_{t-2}	0.1	-	-0.22	-
T^{below}_{t-1}	0.01	0.55	0.07	0.25
T^{below}_{t-2}	-0.05	-0.22	-0.05	0.034
T^{above}_{t-1}	0.02	-0.21	0.12	0.031
T^{above}_{t-2}	-0.02	-0.15	0.01	0.19
T^{off}_{t-1}	0.19	-0.51	-	-
T^{off}_{t-2}	-0.23	0.14	-	-
T^{cor}_{t-1}	0.12	1.00	0.08	0.19
T^{cor}_{t-2}	-0.07	0.75	-0.09	0.03
T^{ext}_{t-1}	0.02	0.11	0.03	0.08
T^{ext}_{t-2}	0.02	0.21	0	0.18

TABLE 4.2 – Coefficients des modèles entrée-sortie en été

Comparaison des différents modèles :

Nous avons tracé sur figures III.12 et III.13 les sorties des modèles $S1$ et $S2$ et la température T^{int} mesurée afin de comparer les différents modèles. Si les données étaient plus importantes on pourrait tester ces modèles sur un autre échantillon de températures de mois correspondant à celui de la construction par exemple.
Les modèles ne comportant pas T^{off} (figure III.13) captent moins bien les changements brefs de comportement. On peut remarquer que T^{int} et T^{off} ont le même comportement (figure : III.16) ce qui explique que les premiers modèles (figure III.12) s'ajustent mieux à T^{int} donnée. Garder T^{off} en entrée cache l'influence des autres variables sur T^{int}.
La régression permet de mieux capter les instants où la courbe s'aplatit par exemple, celle-ci ne faisant intervenir que quelques instants passés. Le meilleur modèle en prédiction semble être $S1 - B$.

Remarque 12. *Les co-linéarités dans un modèle rendent celui-ci bien souvent meilleur en prédiction. Cependant si l'on souhaite étudier la sensibilité, ces modèles ne semblent pas adaptés. Les matrices du modèle peuvent être mal estimées ce qui rend les interprétations douteuses.*

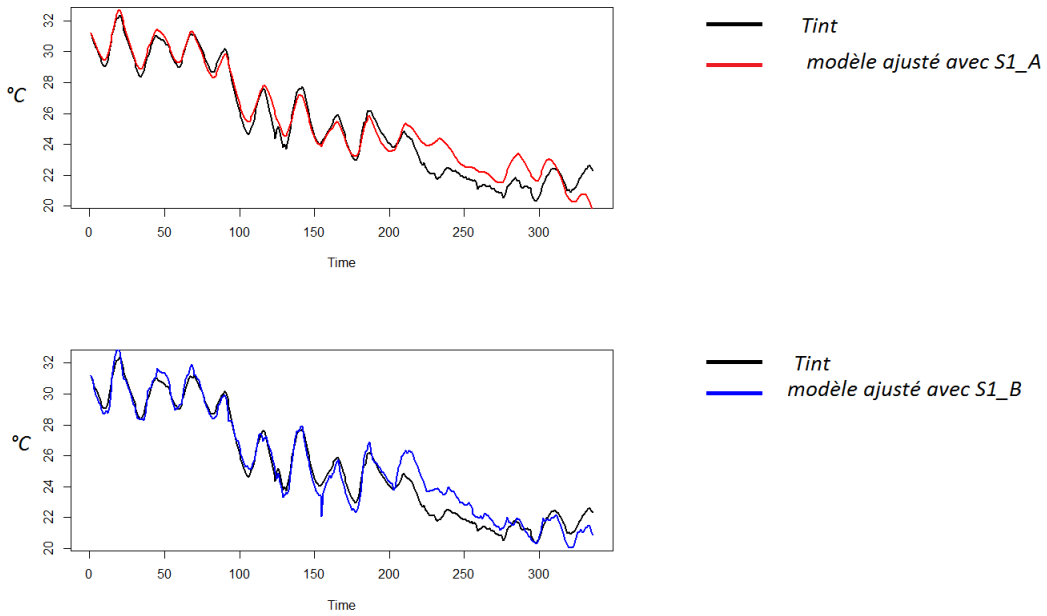

FIGURE III.12 – Comparaison de la sortie T^{int} et des différents modèles $S1 - A$ $(VARX)$ et $S1 - B$ (régression). Vecteur d'entrée $U_t = (T^{\text{below}}, T^{\text{above}}, T^{\text{cor}}, T^{\text{off}}, T^{\text{ext}})$

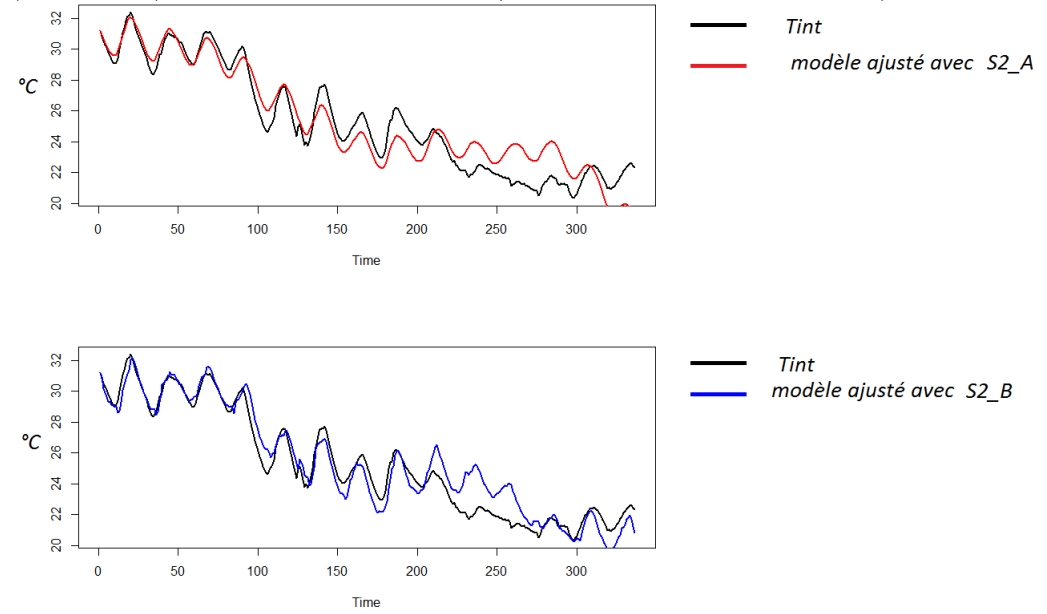

FIGURE III.13 – Comparaison de la sortie T^{int} et des différents modèles $S2 - A$ $(VARX)$ et $S2 - B$ (régression). Vecteur d'entrée $U_t = (T^{\text{below}}, T^{\text{above}}, T^{\text{cor}}, T^{\text{ext}})$

4.2.2 Modèle hiver : K

Nous allons maintenant appliquer ces outils de modélisation pour le modèle hiver ayant comme sortie le chauffage et en entrée l'équivalant chaleur lié aux occupants et les différentes températures $U'_t = (T^{\text{below}}, T^{\text{above}}, T^{\text{off}}, T^{\text{cor}}, T^{\text{ext}}, T^{\text{int}}, N)$.

Le chauffage fonctionne entre 8h et 18h et il est coupé la nuit. Les deux premières semaines, il est coupé le week-end (figure : III.14). Il y a donc deux parties dans cet échantillon qui n'ont pas la même saisonnalité. Garder l'ensemble complet ne permet pas d'estimer les bonnes saisonnalités. Pour faire le pré-traitement, nous allons donc diviser en deux ensembles les données. Une fois ces deux ensembles pré-traités, on les réunira pour obtenir la trajectoire complète d'un processus réduit désigné par K_t.

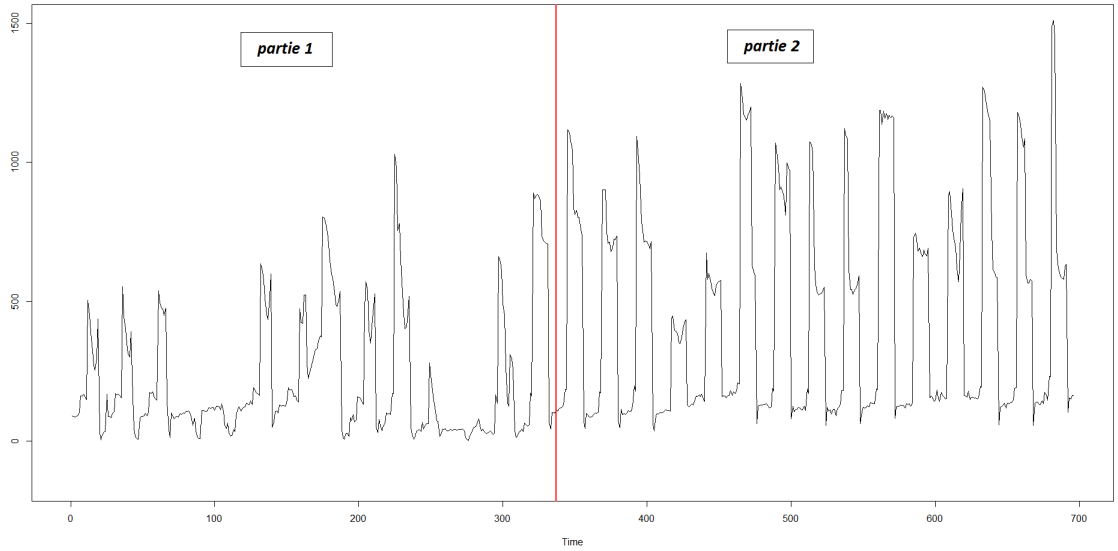

FIGURE III.14 – Chauffage en fonction de l'heure. Mise en évidence du découpage du chauffage en deux parties

Dans ce dernier modèle, nous proposons d'étudier en sortie la variable de chauffage en fonction des différentes températures des pièces environnantes. Cette application peut aider à trouver les pièces ou les apports qui influencent le plus cette dépense énergétique. Il n'y a plus de problème de forte corrélation entre les variables. Nous ne fusionnerons donc plus les températures T^{int} et T^{off}. Nous considérons un modèle entrée-sortie VAR cyclo-stationnaire. Le chauffage et l'équivalent chaleur, notamment, sont des processus présentant de fortes périodicités. Leurs variations pendant la journée sont plus importantes que pendant la nuit. Si l'on considère un modèle qui ne différencie pas ces deux périodes, les variations nocturnes ne seront pas prises en compte car elles seront masquées par celles de la journée. Le processus entre 8h et 18h ne sera pas le même que celui associé à la période de 18h-8h. De plus, intuitivement, on comprend bien que par exemple si on se place à 19 heures, la valeur considérée du chauffage va dépendre de la valeur du chauffage à 18 heures. Entre ces deux heures la puissance de chauffage a chuté. Le coefficient liant ces deux instants ne peut pas être le même que pendant toute la période nuit où le chauffage à un comportement plus "constant". On ne peut donc se placer dans le cas d'un $VAR(1)$ classique. Pour prendre en compte ces variations périodiques nous avons construit alors un modèle cyclo-stationnaire, c'est-à-dire que les coefficients de la matrice de régression dépendent de l'instant t considéré :

$$K_t = a_t K_{t-1} + \boldsymbol{B}_t \boldsymbol{U}'_{t-1} \tag{III.6}$$

Le coefficient a_t et le vecteur $\boldsymbol{B}_t$ dépendent de l'instant t et sont périodiques de période P, ici $P = 24$, c'est-à-dire :

$$a_{t+kP} = a_t$$
$$\boldsymbol{B}_{t+kP} = \boldsymbol{B}_t$$

Nous avons choisi de ne pas prendre en compte les coupures des week-ends. La période est alors de 24h. Si l'on avait choisi d'en tenir compte, la période aurait été de 7 jours. Dans notre échantillon seulement deux week-ends apparaissent. Le peu de données et le nombre de paramètres à estimer plus important dans ce cas, nous a poussé à ne pas en tenir compte.

Les coefficients vont être estimés sous la forme :

$$a_t = a + b\cos(\frac{2\pi t}{P}) + c\sin(\frac{2\pi t}{P}) \tag{III.7}$$
$$\boldsymbol{B}_t = \boldsymbol{A} + \boldsymbol{B}\cos(\frac{2\pi t}{P}) + \boldsymbol{C}\sin(\frac{2\pi t}{P}) \tag{III.8}$$

L'équation III.6 peut se réécrire alors sous la forme :

$$K_t = (a + b\cos(\frac{2\pi t}{P}) + c\sin(\frac{2\pi t}{P}))K_{t-1} + (\boldsymbol{A} + \boldsymbol{B}\cos(\frac{2\pi t}{P}) + \boldsymbol{C}\sin(\frac{2\pi t}{P}))U'_{t-1} \tag{III.9}$$

En posant : $K^c_{t-1} = \cos(\frac{2\pi t}{P})K_{t-1}$, $K^s_{t-1} = \sin(\frac{2\pi t}{P})K_{t-1}$ et
$\boldsymbol{U}'^c_{t-1} = \cos(\frac{2\pi t}{P})\boldsymbol{U}'_{t-1}$, $\boldsymbol{U}'^s_{t-1} = \sin(\frac{2\pi t}{P})\boldsymbol{U}'_{t-1}$, on obtient le modèle suivant :

$$K_t = aK_{t-1} + bK^c_{t-1} + cK^s_{t-1} + \boldsymbol{A}\boldsymbol{U}'_{t-1} + \boldsymbol{B}\boldsymbol{U}'^c_{t-1} + \boldsymbol{C}\boldsymbol{U}'^s_{t-1} \tag{III.10}$$

Nous estimons $(a, b, c, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C})$ par maximum de vraisemblance puis nous obtenons les coefficients de a_t et $\boldsymbol{B}_t$ à partir de III.7.

En figure III.15, nous avons tracé le chauffage K mesurée en noir et celui obtenu à partir du modèle. Le modèle semble bon. Cependant ce modèle ne capte peut être pas suffisamment les faibles variations de la nuit. Il n'est peut être pas suffisamment compliqué ou estimé de manière assez fine. En effet la faible quantité de données et leur piètre qualité (concaténation de deux périodes différentes) ne permettent pas d'obtenir un modèle idéal.

4.3 Les modèles d'entrées

4.3.1 Pré-traitement, variables réduites

La méthode d'analyse de sensibilité proposée repose sur la forme du modèle d'entrée. Du fait du nombre insuffisant de données disponibles pour appliquer la méthode Pick and Freeze, nous proposons des modèles de simulation des entrées. Ces variables dépendant du temps, seront modélisées par des séries temporelles. Le plus simple pour construire un modèle de

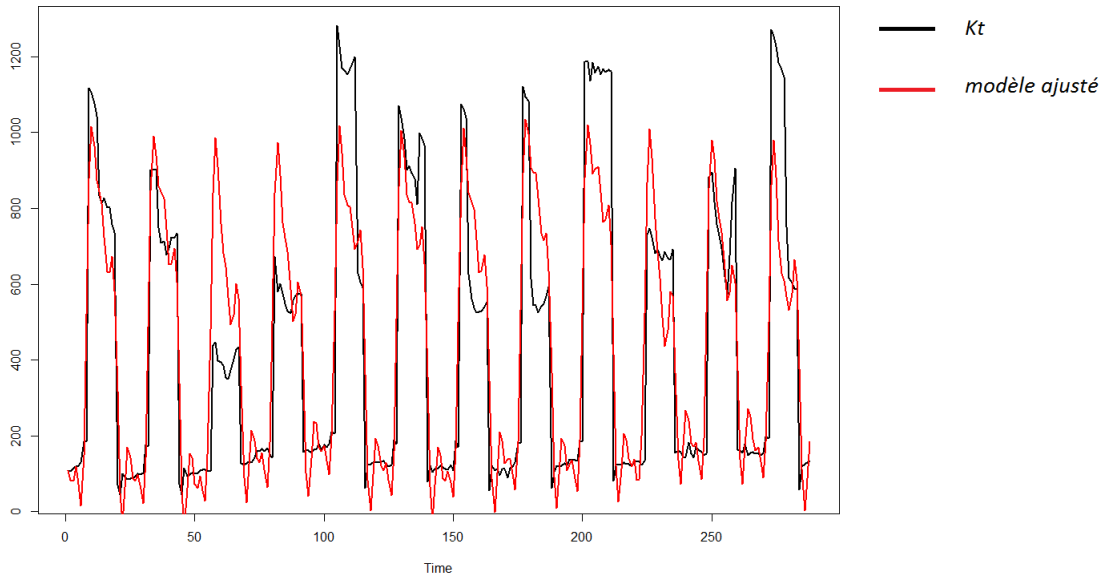

FIGURE III.15 – Comparaison de la sortie K et du modèle ajusté

simulation efficace est d'essayer de se ramener à des séries stationnaires ou cyclo-stationnaires. La plupart du temps il suffit d'enlever les variations lentes déterministes, appelées tendances pouvant être additives (m_t^Z) et/ou multiplicatives (h_t^Z). A celles ci peuvent s'ajouter un effet périodique (saison) notée s_t^Z (additive) et v_t^Z (multiplicative). Les fonctions s_t^Z et v_t^Z sont donc des fonctions périodiques. Dans notre cas les séries présentent une forte saisonnalité additive (notée s_t) et/ou multiplicative (notée v_t) de 24 heures. Il est possible que le processus possède plusieurs saisonnalités. Par exemple le chauffage en 2009 possède une saisonnalité de 24 heures et une saisonnalité de 7 jours.
Chaque variable est décomposée alors de la manière suivante :

$$Z_t = s_t^Z + m_t^Z + v_t^Z h_t^Z Z_t^r$$

Les saisonnalités et tendances ont été estimées par les méthodes expliquées dans le chapitre 1 de la partie 2.

Remarque 13. *Du point de vue technique, si la modélisation est cyclo-stationnaire la tendance saisonnière sur la variance est un des éléments de la saisonnalité de la covariance et donc on ne peut utiliser des processus normés pour effectuer l'analyse de sensibilité. Si les processus ont besoin d'être normés, il faudra re-multiplier les variables par* $v_t^Z h_t^Z$.

Z_t^r appelé processus réduit, s'obtient en retirant les différentes saisonnalités et tendances. Ce processus est alors centré et normé.

Le processus multi-dimensionnel $(Z_t^1, \ldots, Z_t^p)$ sera modélisé ensuite par un processus $VAR(1)$ (stationnaire ou cyclo-stationnaire) :

$$Z_t^r = AZ_{t-1}^r + \epsilon_t$$

ou par un processus cyclo-stationnaire VAR de période P :

$$Z_t^r = A_{[t]} Z_{t-1}^r + \epsilon_t$$

où $[t] = t \bmod P$.
La matrice A ou $A_{[t]}$ est estimée par les méthodes proposées dans le chapitre 1 de la partie 2. Une fois estimée, les modèles sont validés par les opérations suivantes :

- Par bootstrap des résidus estimés $\epsilon_t^* = Z_t^r - \hat{A} Z_{t-1}^r$, on estime des intervalles de confiance pour les éléments de A (pour avoir une idée de l'incertitude sur les paramètres des processus d'entrées).
- On teste la blancheur de la suite des (ϵ_t^*) et son caractère gaussien en utilisant des tests classiques (test du Portemanteau par exemple).

4.3.2 Les modèles été : E-A et E-B

On commence par les modèles été qui ne contiennent que des températures (il n'y a pas de chauffage ni de données sur les occupants).
Elles sont dessinées sur la figure : III.16.

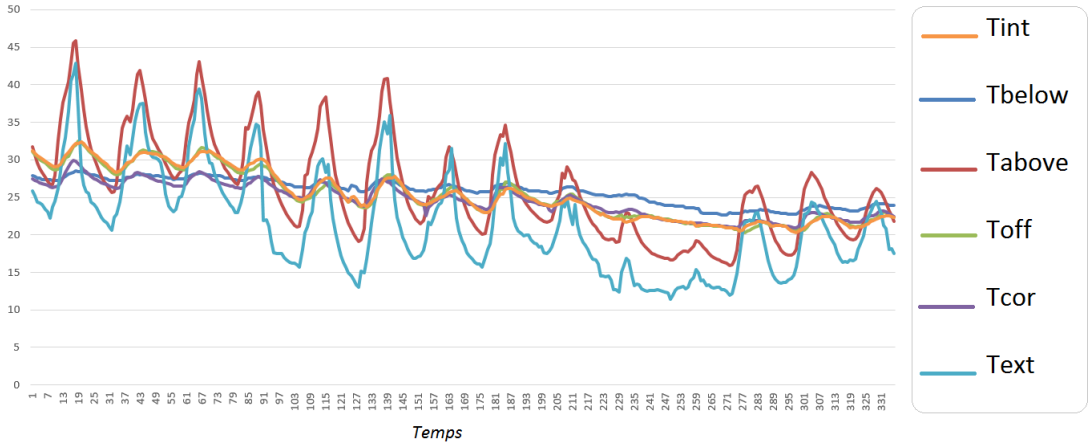

FIGURE III.16 – Différentes températures mesurées en fonction du temps du 22/08/2012 au 04/09/2012

Toutes les températures des différentes pièces sont plus ou moins corrélées entre elles (voir figure : III.17). Globalement toutes les corrélations sont très importantes (table 4.3). Nous faisons le choix d'un modèle où le vecteur $\boldsymbol{U}_t$ contient toutes les températures :

$$\boldsymbol{U}_t = (T_t^{\text{below}}, T_t^{\text{above}}, T_t^{\text{off}}, T_t^{\text{cor}}, T_t^{\text{ext}})$$

L'idée première était de découper le bâtiment en un maximum de pièces afin de savoir au mieux laquelle influençait la température intérieure. Le problème de cette démarche peut impliquer la sur-paramétrisation du modèle et des confusions d'effet. Par exemple dans notre cas, la température du bureau est très proche de la température de la pièce étudiée. On remarque

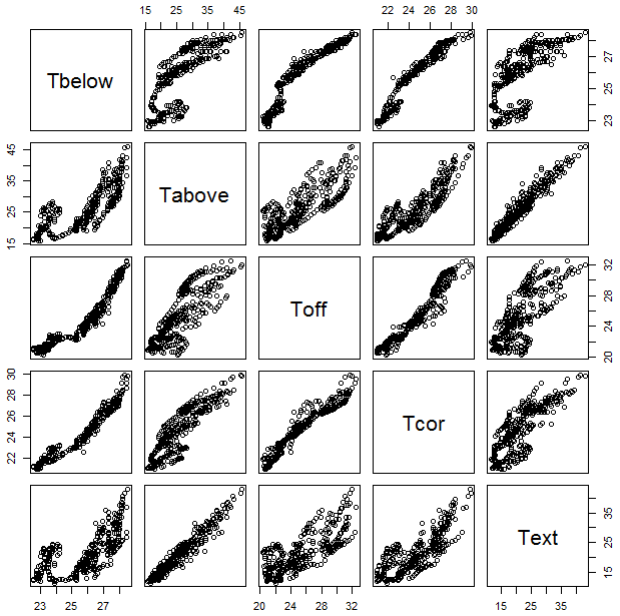

FIGURE III.17 – Scatterplot des différentes températures

	T_t^{below}	T_t^{above}	T_t^{off}	T_t^{cor}	T_t^{ext}	T_t^{int}
T_t^{below}	1	0.38	0.66	0.64	0.36	0.64
T_t^{above}	0.38	1	0.56	0.77	0.88	0.62
T_t^{off}	0.66	0.55	1	0.70	0.58	0.92
T_t^{cor}	0.64	0.77	0.70	1	0.72	0.75
T_t^{ext}	0.36	0.88	0.58	0.72	1	0.61
T_t^{int}	0.64	0.62	0.92	0.75	0.61	1

TABLE 4.3 – Matrice de corrélation des différentes températures

une forte corrélation entre $(T^{\text{off}}, T^{\text{int}})$ (0.92) qui avait déjà été évoquée lors du choix des modèles été de sortie. Une corrélation de 0.92 signifie en fait que ces deux variables sont quasi co-linéaires. Inclure T^{off} et T^{int}, en entrée respectivement en sortie, peut masquer le rôle des autres variables. Le fait de garder T^{off} dans le vecteur d'entrée dans l'étude peut dissimuler d'autres interactions. Dans un deuxième modèle, nous suggérons de réunir le bureau et la pièce d'intérêt en une seule pièce d'un point de vue thermique.
Le nouveau $\boldsymbol{U}_t$, noté $\boldsymbol{U}_t^1$ est alors :

$$\boldsymbol{U}_t^1 = (T_t^{\text{below}}, T_t^{\text{above}}, T_t^{\text{cor}}, T_t^{\text{ext}})$$

Pour résumer, on peut envisager principalement deux modèles :

1. **modèle E-A** : où le processus aléatoire d'entrée considéré est : $\boldsymbol{U}_t$
2. **modèle E-B** : en remarquant la trop forte corrélation (0.92) du couple $(T^{\text{off}}, T^{\text{int}})$ on considère un modèle qui ne prend plus en compte en entrée T^{off} (physiquement T^{int}

et T^{off} ne formeront plus qu'une seule zone thermique). Le processus aléatoire d'entrée considéré est donc U_t^1

E-A multivarié On considère le processus vectoriel $\boldsymbol{U}_t$ donné par $\boldsymbol{U}_t = (T^{\text{below}}, T^{\text{above}}, T^{\text{off}}, T^{\text{cor}}, T^{\text{ext}})$

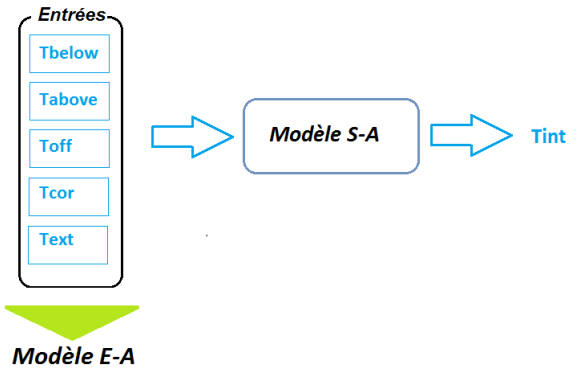

FIGURE III.18 – Schématisation du modèle **E-A**. Modèle **S-A**, modèle entrée-sortie correspondant au modèle d'entrée **E-A**

Les coefficients du modèle sont estimés à partir de la log-vraisemblance pénalisée (voir 1 partie 2) :

$$\boldsymbol{U}_t = \sum_{k=1}^{p} A_k \boldsymbol{U}_{t-k} + \epsilon_t \sim N(0, \sigma^2)$$

Remarque 14. *La méthode d'Akaike ne teste que les ordres globaux, c'est-à-dire des matrices A_k carrées. Tester tous les sous modèles rectangulaires prend un temps considérable. Nous avons remarqué que si on annule tous les coefficients inférieurs à 0.05 on obtient une matrice A_k beaucoup plus creuse avec un modèle meilleur suivant le critère d'Akaike que la matrice initiale.*

Nous n'avons pas appliqué de méthodes de recherche de "sparsity" sur les paramètres (type Lasso) qui sont peut être une issue si la dimension p est beaucoup plus grande

Le modèle estimé est un VAR de mémoire $p = 2$. Pour faciliter les simulations des entrées on peut remettre les processus $VAR(p)$ sous forme d'un processus $VAR(1)$.

En effet on peut remarquer que tous les processus $VAR(p)$ peuvent se réécrire par un processus $VAR(1)$:

$$T_t = \sum_{l=1}^{p} d_l T_{t-p} + \omega_t$$

Pour chaque p fixé on estime $(d_l, \quad l \leq 1, \ldots, p)$ et Θ la matrice de covariance de ω par maximum de vraisemblance pour T_t stationnaire. p est choisi en utilisant un critère AIC.

On pose

$$\boldsymbol{Z}_t = (T_t, T_{t-1}, \ldots, T_{t-p})^* = A\boldsymbol{Z}_{t-1} + \xi_t$$

où A est une matrice $(p \times p)$ telle que :

$$A = \begin{pmatrix} d_1 & d_2 & d_3 & \cdots & d_p \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}, \quad \xi_t = \begin{pmatrix} \omega_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{III.11}$$

Dans notre cas le vecteur $\boldsymbol{Z}_t$ s'écrit : $\boldsymbol{Z}_t = (T_t^{\text{below}}, T_{t-1}^{\text{below}}, T_t^{\text{above}}, T_{t-1}^{\text{above}}, T_t^{\text{off}}, T_{t-1}^{\text{off}}, T_t^{\text{cor}}, T_{t-1}^{\text{cor}}, T_t^{\text{ext}}, T_{t-1}^{\text{ext}})$
Les coefficient d_l de la matrice A sont transformés en matrice bloc de taille (2×2). A sera donc de taille $((2 * 5) \times (2 * 5))$. Les matrices de covariance seront elles aussi recomposées.

	T_t^{below}	T_t^{above}	T_t^{off}	T_t^{cor}	T_t^{ext}
T_{t-1}^{below}	0.88	-0.23	0.21	0.12	0.20
T_{t-2}^{below}	0.04	0.03	-0.16	-0.03	-0.45
T_{t-1}^{above}	0.01	1.21	0.03	0.06	0.61
T_{t-2}^{above}	-0.01	-0.48	-0.02	-0.04	-0.60
T_{t-1}^{off}	0.06	0.01	1.24	0.03	-0.09
T_{t-2}^{off}	-0.07	-0.04	-0.3	-0.05	0.25
T_{t-1}^{cor}	0.06	0.49	0.06	1	0.71
T_{t-2}^{cor}	-0.04	-0.08	-0.09	-0.15	-0.31
T_{t-1}^{ext}	0	0.3	0.01	0.01	0.85
T_{t-2}^{ext}	0.01	-0.04	-0.1	-0.01	-0.05

TABLE 4.4 – Coefficients du **modèle E-A**

Le même travail est effectué pour le vecteur $\boldsymbol{U}_t^1$. Les tableaux des coefficients de A et des matrices de covariances sont présentés en annexe.

Comparaison des modèles Pour choisir lequel des deux modèles est meilleur en prédiction, nous calculons ensuite la différence au carré entre la prévision et les températures réduites (voir table 4.5). Les deux modèles donnent des résultats similaires cependant le modèle $E - B$ est plus simple car il comporte moins de variable.

4.3.3 Modèle hiver des entrées

Ce modèle d'entrée proposé correspond au modèle de sortie du chauffage III.6. En hiver s'ajoutent deux variables d'entrées : la température interne qui devient une entrée et l'équi-

	T_t^{below}	T_t^{above}	T_t^{off}	T_t^{cor}	T_t^{ext}
E-A	4	209	10	9	452
E-B	4	208	-	9	453

TABLE 4.5 – Comparaison des modèles été

valent chaleur des occupants N_t. Le vecteur d'entrée considéré devient :

$$\boldsymbol{U}'_t = (T_t^{\text{below}}, T_t^{\text{above}}, T_t^{\text{off}}, T_t^{\text{cor}}, T_t^{\text{ext}}, T_t^{\text{int}}, N_t)$$

Les personnes ne sont présentes que durant les jours-ouvrés de 8h à 18h. Cet aspect cyclique nous oblige à distinguer ces différentes périodes. Si l'on considère un modèle qui ne différencie pas ces deux périodes, les changements de régimes à 8h et 18h ne seront pas pris en compte. De plus, en raison de l'absence de personnes la nuit on comprend bien que le processus entre 8h et 18h ne sera pas le même que celui associé à la période de 18h-8h.

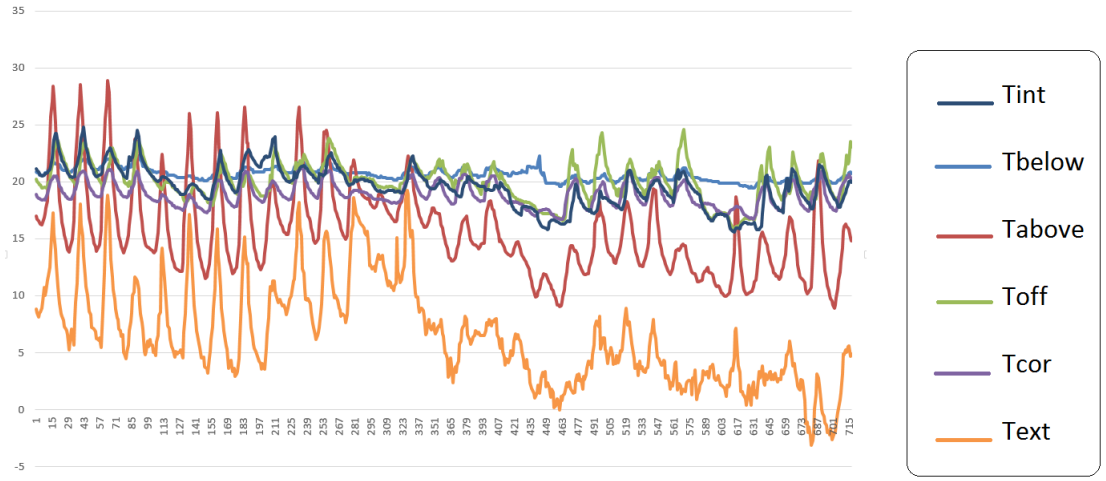

FIGURE III.19 – Différentes températures en fonction du temps du 13/11/2012 au 13/12/2012

Les températures seront normées et centrées. Nous avons décidé de normer les données dans ce cas ci. En effet le vecteur des entrées comporte à la fois des températures exprimées en degrés et des flux de chaleur exprimés en kilos watt. Les températures n'auront pas les mêmes variances que le flux de chaleurs associé aux personnes. Pour avoir une homogénéité des variations nous les normons.

Pour le flux, nous éliminons les données de nuit et de week-ends avant prétraitement dont la valeur nulle modifierait les moyennes et variances (figure : III.20). S'il y a une saisonnalité sur les profils journaliers et sur les cinq jours de la semaine, elle est déterminée puis éliminée pour travailler sur la variable réduite. Le flux une fois normé est reconstitué en ajoutant des valeurs nulles correspondant aux nuits et week-ends.

Nous avons construit un modèle multivarié cyclo-stationnaire de période 24h du type :

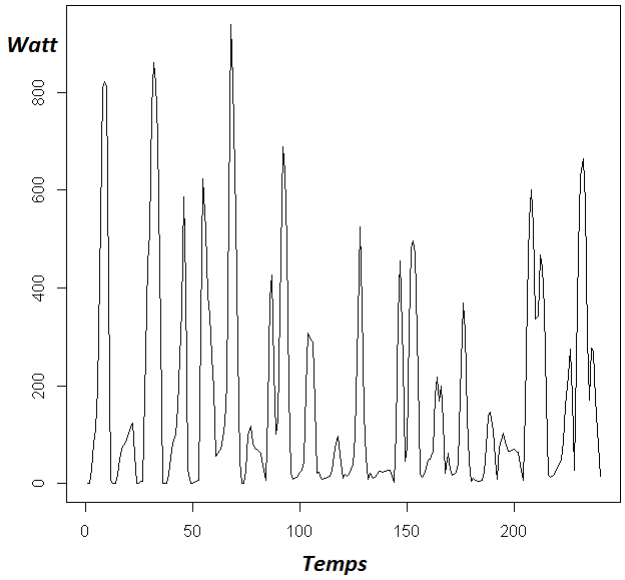

FIGURE III.20 – Équivalent chaleur des occupants de 8h à 18h les jours ouvrés du 13/11/2012 au 12/12/2012

$$\boldsymbol{U}_t = \boldsymbol{A}_t \boldsymbol{U}_{t-1} \tag{III.12}$$

La matrice $\boldsymbol{A}_t$ dépend de l'instant t et est périodique de période $P = 24h$ c'est-à-dire :

$$\boldsymbol{A}_{t+kP} = \boldsymbol{A}_t \tag{III.13}$$

Les coefficients vont être estimés sous la forme :

$$\boldsymbol{A}_t = \boldsymbol{A}^1 + \boldsymbol{A}^2 H_t \tag{III.14}$$

où H_t est une fonction de type créneaux de période 24h (modèle avec switch) :

$$\begin{cases} H(t) & = 0, \text{ pour } 0 \leq t \leq 6 \text{ et } 18 \leq t \leq 23 \\ H(t) & = 1, \quad 9 \leq t \leq 16 \\ H(t) & = \dfrac{1}{2}t - 3.5, \quad 6 \leq t \leq 8 \\ H(t) & = \dfrac{-1}{2}t + 9.5, \quad 16 \leq t \leq 23 \end{cases} \tag{III.15}$$

Cette fonction plus simple que le fait de caler une fonction sinus et cosinus (voir modèle entrée-sortie 4.2.2) permet d'estimer moins de paramètres et donc un modèle de meilleure qualité au vu du peu de données disponibles.

Nous reconstituons ensuite la matrice périodique de la même manière que pour le modèle entrée-sortie concernant le modèle de chauffage.

Les coefficients sont estimés par maximum de vraisemblance ainsi que la matrice de covariance.

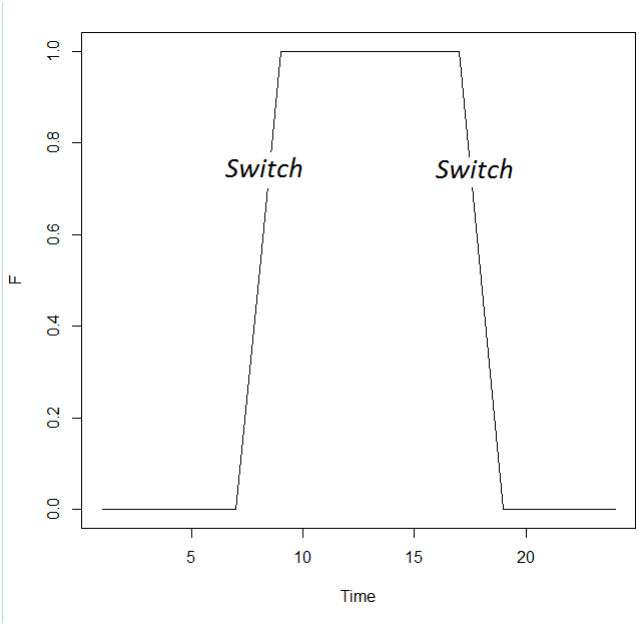

FIGURE III.21 – Fonction H

La matrice de covariance sera aussi périodique. On imposera pour les périodes correspondant à la nuit, un bruit nul pour les lignes et colonnes correspondant à l'équivalent chaleur du nombre de personnes, cela afin de garantir que le modèle retranscrive le fait qu'il n'y ait pas de visiteurs la nuit.

Chapitre 5

Analyse de sensibilité

5.1 Méthode

On souhaite calculer l'indice de sensibilité associé à ces modèles à partir d'une méthode Pick and Freeze. Cette méthode nécessite deux échantillons d'entrée $SIMU_1$ et $SIMU_2$ où toutes les quantités correspondant à $\boldsymbol{U}^i$ (la variable d'entrée i) sont gelées. Les sorties respectives correspondant à ces entrées sont calculés (Y_1 et Y_2) puis des estimateurs empiriques des variances conditionnelles et des variances sont calculées afin d'obtenir un estimateur de l'indice de sensibilité correspondant à l'entrée $\boldsymbol{U}^i$.

Algorithm 2 Méthode d'estimation Pick and Freeze

Require: $SIMU_1, SIMU_2, init$

$times \leftarrow dim(SMU_1)[1]$, $input \leftarrow dim(SIMU_1)[2]$
$indice \leftarrow 0$
for $i = 0$ to $input$ **do**
  if init $\neq$ NULL **then**
    $Y_1 = f(SIMU_1, init)$
    $Y_2 = f(SIMU_2, init)$
  else
    $Y_1 = f(SIMU_1)$
    $Y_2 = f(SIMU_2)$
  end if
  $indice[, input] = \dfrac{\mathbf{E}(Y_1 \otimes Y_2) - \mathbf{E}(Y_1)\mathbf{E}(Y_2)}{\mathbf{Var}(Y_1)}$
end for
return $indice$

Pour obtenir les deux échantillons d'entrée ($SIMU_1$ et $SIMU_2$) de taille N dans le cas d'entrées corrélées comme c'est le cas ici, nous avons utilisé la méthode de séparation des variables présentée dans le chapitre précédent (2.2). Deux algorithmes sont présentés : un pour le cas stationnaire (3) et un pour le cas cyclo-stationnaire (4). L'algorithme pour le cas cyclo-stationnaire peut s'adapter au cas quelconque qui correspondant à une période $P = \infty$. Les

modèles de simulation des processus utilisés sont des modèles auto-régressifs. Cependant nous pouvons utiliser d'autres types de modèles de simulation de type Gaussien sans changer de méthode.
Ces algorithmes sont ensuite appliqués aux modèles précédemment décrits.

5.1.1 Cas stationnaire

Algorithm 3 Réduction à des entrées indépendantes : cas stationnaire

Require: $A, \Theta, U, N, init_1, init_2$
1: $times \leftarrow dim(U)[1]$, $input \leftarrow dim(U)[2]$
2: $Simu_1[times, input, N] \leftarrow 0$, $Simu_2[times, input, N] \leftarrow 0$
3: **for** $i = 1$ to $input$ **do**
4: $\omega_1[times, input, N] \sim \mathcal{N}(0, \Theta)$
5: $\omega_2[times, input, N] \sim \mathcal{N}(0, \Theta)$
6:
7: $Simu_1[1, ,] \leftarrow init_1$
8: $Simu_2[1, ,] \leftarrow init_2$
9: **for** $t = 2$ to $times$ **do**
10: $Simu_1[t, ,] \leftarrow A \cdot Simu_1[t-1, ,] + \omega_1[t, ,]$
11: $Simu_2[t, ,] \leftarrow A \cdot Simu_2[t-1, ,] + \omega_2[t, ,]$
12: **end for**
13:
14: **for** $t = 1$ to $times$ **do**
15: $\Lambda \leftarrow (\mathbf{Cov}(U[1:t, i], U[1:t, i]))^{-1}\mathbf{Cov}(U[1:t,], U[1:t,])$
16: $\tilde{X}_1[t, ,] \leftarrow (Simu_1[1:t, i,])^* \cdot \Lambda$
17: $\tilde{X}_2[t, ,] \leftarrow (Simu_2[1:t, i,])^* \cdot \Lambda$
18: **end for**
19: $W_1 \leftarrow Simu_1 - \tilde{X}_1$
20: $W_2 \leftarrow Simu_2 - \tilde{X}_2$
21: $SIMU_1 \leftarrow Simu_1$
22: $SIMU_2 \leftarrow \tilde{X}_1 + W_2$
23: **end for**
24: **return** $(SIMU_1, SIMU_2)$

Le premier algorithme proposé s'applique à des données stationnaires. Les données simuler dans cet algorithme sont des processus VAR (Etapes 4 à 12). On aurait pu utiliser n'importe quel autre processus stationnaire.

Après avoir simulé deux échantillons : $SIMU_1$ et $SIMU_2$ sur la totalité de la période de temps choisies ($times$), on applique la méthode de séparation des variables. Le fait de remplacer par $W_1 \leftarrow Simu_1 - \tilde{X}_1$, toute la partie jusqu'à $times$, pour chaque entrée, sera gelée. On calculera alors les indices :

$$S_{1,1}, S_{2,2}, \ldots, S_{times,times}$$

L'avantage d'utiliser des données stationnaires repose sur le fait que l'indice que l'on calcule est indépendant de l'instant : $S_{t,k} = S_k$ pour tout k.
On aura alors toutes les mémoires à tous les instants.

La période *times* choisie dépend de la vitesse de convergence de l'indice. Le calcul de Λ étant lourd, on a tout intérêt à choisir *times* le plus petit possible afin d'optimiser le temps de calcul.

Comme nous utilisons des processus stationnaires, nous devons prendre garde aux valeurs initiales utilisées pour être assuré que les processus $SIMU_1$ et $SIMU_2$ le soient. Pour cela, on peut utiliser un troisième processus $SIMU_{init}$, le simuler avec une initialisation quelconque et sélectionner deux valeurs de temps t_1 et t_2 éloignées. Ces valeurs seront $init_1$ et $init_2$ utilisées aux étapes 7 et 8 pour initialiser $SIMU_1$ et $SIMU_2$. En général t_1 et t_2 sont choisis aux alentours de 1000. Si les données ne sont pas stationnaires on pourra observer que l'indice n'est pas strictement croissant en k.

Pour pouvoir calculer les indices, si les entrées ont été normées, il faut à ce stade multiplier $SIMU_1$ et $SIMU_2$ par les saisonnalités et tendances multiplicatives correspondantes aux entrées.

5.1.2 Cas cyclo-stationnaire

Dans cet algorithme on se place dans un cas cyclo-stationnaire. Il est comme précédemment présenté pour des processus VAR cyclo-stationnaire (Etapes 7 à 12). On aurait pu utiliser n'importe quel autre processus cyclo-stationnaire à la place.

L'entrée étant un processus cyclo-stationnaire de période P, l'indice de sensibilité $S_{t,k}$ sera périodique :

$$S_{t,k} = S_{t+hP,k} \text{ pour tous } h$$

Cette propriété de périodicité réduit considérablement les calculs en permettant de ne calculer que P indices. On va calculer pour chaque $t \in [0, P]$ fixé, $S_{t,k}$ de la même manière que précédemment. A chaque instant, pour le même processus $SIMU_1$ et $SIMU_2$, on applique successivement la méthode précédente (3) pour le processus décalé de t_1 à *times* avec $t_1 \in [\![0, P]$. W_1 est remplacer par $W_1 \leftarrow Simu_1 - \tilde{X}_1$ jusqu'à *times*, pour chaque entrée. On calculera alors à chaque itération les indices résumés dans la table : 5.1.

L'indice pour t fixé se lit à la ligne P du tableau : 5.1 et de droite à gauche. On peut noter alors que *times* doit être au minimum être égal à deux fois la période pour pouvoir obtenir tous les indices. Donc pour chaque variable, P processus sont simulés. Ainsi nous aurons tous les indices désirés (table : 5.1).

Si l'on souhaite étudier l'indice à une mémoire fixée k, les indices sont obtenus dans la diagonale de la table : 5.1.

De la même manière que dans le cas stationnaire, des processus sont simulés afin de permettre une initialisation des processus $Simu_1$ et $Simu_2$ de manière à ce qu'ils soient stationnaires. Il faut cependant prendre garde aux instants utilisés pour initialiser ces processus qui doivent correspondre à l'instant $t = 0$ c'est-à-dire tous multiple de P.

On pourrait adapter cet algorithme à un processus quelconque en prenant une période égale à $P = \infty$. On peut choisir pour l'appliquer une période P grande de manière arbitraire.

itération, P= / t_1	1	2	3	...	P
1	$S_{1,1}$	-	-	-	-
2	$S_{2,2}$	$S_{2,1}$	-	...	-
3	$S_{3,3}$	$S_{3,2}$	$S_{3,1}$	$\ddots$	-
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
P	$S_{P,P}$	$S_{P,(P-1)}$	$S_{P,(P-2)}$	...	$S_{P,1}$
P+1	$S_{P+1,P+1}$	$S_{P+1,P}$	$S_{P+1,(P-1)}$	...	$S_{P+1,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2*P	$S_{2P,2P}$	$S_{2P,2P-1}$	$S_{2P,2P-2}$	...	$S_{2P,P}$

TABLE 5.1 – Tableau récapitulatif du calcul des indices dans le cas cyclo-stationnaire

Algorithm 4 Réduction à des entrées indépendantes : cas cyclo-stationnaire

Require: $A, \Theta, U, n, Simu_1, Simu_2, init_1, init_2$

$times \leftarrow dim(U)[1]$, $input \leftarrow dim(U)[2]$
$Simu_1[times, input, n] \leftarrow 0$, $Simu_2[times, input, n] \leftarrow 0$, $Simu_{init}[times, input, n] \leftarrow 0$
for $i = 1$ to $input$ **do**
  $\omega_1[times, input, n] \sim \mathcal{N}(0, \Theta)$
  $\omega_2[times, input, n] \sim \mathcal{N}(0, \Theta)$

  $Simu_1[1, ,] \leftarrow init_1$
  $Simu_2[1, ,] \leftarrow init_2$
  for $t = 2$ to $times$ **do**
    $Simu_1[t, ,] \leftarrow A * Simu_1[t-1, ,] + \omega_1[t, ,]$
    $Simu_2[t, ,] \leftarrow A * Simu_2[t-1, ,] + \omega_2[t, ,]$
  end for

  {Boucle pour tenir compte de la périodicité}
  for $P = 1$ to $Periode$ **do**
    for $t = 1$ to $times$ **do**
      $\Lambda \leftarrow (\mathbf{Cov}(U[P:t, i](U[P:t, i]))^*)^{-1}\mathbf{Cov}(U[P:t,], (U[P:t,])^*)$
      $\tilde{X}_1[t, ,] \leftarrow (Simu_1[P:t, i,])^*\Lambda$
      $\tilde{X}_2[t, ,] \leftarrow (Simu_2[P:t, i,])^*\Lambda$
    end for
    $W_1 \leftarrow Simu_1 - \tilde{X}_1$
    $W_2 \leftarrow Simu_2 - \tilde{X}_2$
    $SIMU_1[, , P] \leftarrow Simu_1$
    $SIMU_2[, , P] \leftarrow \tilde{X}_1 + W_2$
    if $P \neq 1$ **then**
      $Simu_1[1:(P-1), ,] = 0$
      $Simu_2[1:(P-1), ,] = 0$
    end if
  end for
end for

return $(SIMU_1, SIMU_2)$

Pour pouvoir calculer les indices, si les entrées ont été normées, il faut à ce stade multiplier $SIMU_1$ et $SIMU_2$ par les saisonnalités et tendances multiplicatives correspondant aux entrées.

5.2 Résultats de l'analyse de sensibilité

5.2.1 Modèles été

Nous nous intéressons dans un premier temps au modèle de sortie $S1 - A$, qui est un modèle auto-régressif avec variable exogène $VARX(2)$ (équation III.4). Les indices sont estimés avec des échantillons de taille $N = 5000$ et sont représentés sur la figure III.1.
Les températures les plus influentes sont la température extérieure T^{ext} et la température de la pièce du dessus T^{above} (leurs indices sont proches de 1) . Les variances des bruits de ces variables sont les plus importantes A.0.1. Elles sont respectivement de 1.34 et 0.62 (table : A.0.1, alors que pour les autres variables elles sont autour de 0.02. Il semble donc raisonnable qu'elles ressortent comme des variables importantes dans l'analyse de sensibilité.
La troisième variable est T^{cor}. De variance faible, elle a une sensibilité néanmoins importante. On peut voir à travers cette variable que la sensibilité n'est pas qu'une affaire de variance importante. La sensibilité est fonction du modèle entrée-sortie mais aussi des covariances de la variable d'intérêt avec les autres variables.
La pièce la plus "froide" est celle du couloir, elle est donc la pièce permettant de faire diminuer la température. Elle est la plus susceptible d'engendrer des variations de la variable d'intérêt. T^{below} n'est pas du tout influente. Cette pièce correspond à la pièce du dessous, isolée par une dalle, il semble cohérent que cette pièce n'ait pas d'influence sur les variations de T^{int}.

Les mémoires utiles sont grandes de l'ordre de 17h pour $(T^{\text{cor}}, T^{\text{ext}}, T^{\text{off}})$ et 24h pour T^{above}. Cela sous entend que le bâtiment à une grande inertie thermique. La température de la pièce intérieure dépend des températures qu'il faisait 17 heures auparavant.
Les oscillations que l'on observe peuvent provenir d'une tendance basse fréquence non corrigée ou bien d'un effet cyclique perturbateur (par exemple hebdomadaire). Il se peut aussi que ces oscillations proviennent de la taille de l'échantillon utilisé pour estimer les indices. En augmentant la taille N, ces oscillations s'atténuent.

Au vue des fortes corrélations entre T^{off} et T^{int} nous avons décidé de supprimer cette entrée et de nous intéresser aux modèles $S2-A$ et $S2-B$ afin de stabiliser et mieux interpréter les valeurs des indices. Ce modèle revient à fusionner T^{off} et T^{int}. Conserver T^{off} crée des colinéarités qui contribuent à des instabilités numériques. Les matrices du modèle entrée-sortie peuvent être mal estimées ce qui rend les interprétations douteuses.

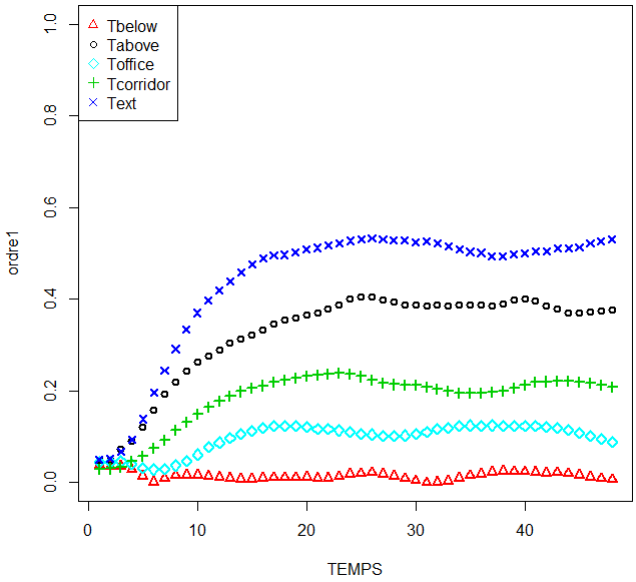

FIGURE III.1 – Indice de sensibilité en fonction du temps. **Modèle E-A** en entrée et entrée-sortie **S1-A** (*VARX*)

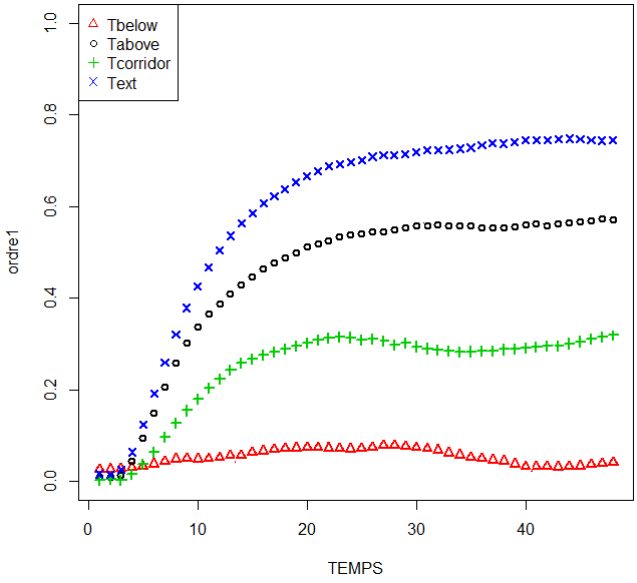

FIGURE III.2 – Indice de sensibilité en fonction du temps. **Modèle E-B** en entrée et modèle de sortie **S2-A** (*VARX*)

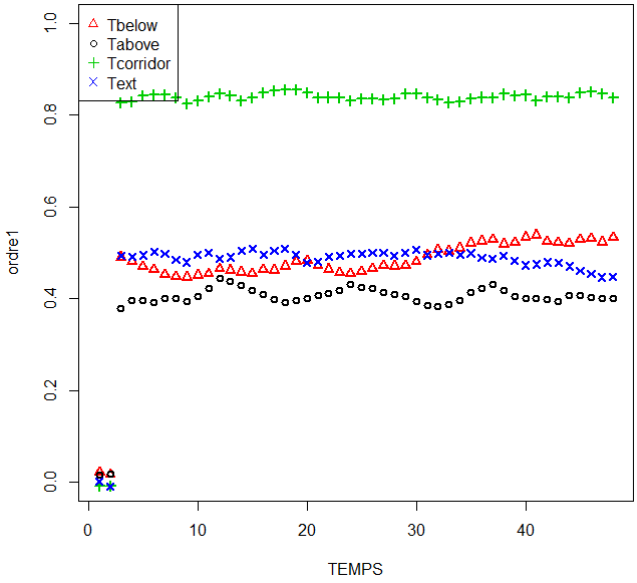

FIGURE III.3 – Indice de sensibilité en fonction du temps. **Modèle E-B** en entrée et modèle de sortie **S2-B** (regression)

Modèle sans T^{off} en entrée : Il apparait sur la figure : III.2 que les températures $(T^{\text{ext}}, T^{\text{above}})$ sont toujours les variables les plus importantes, suivies par T^{cor} et T^{below}. T^{below} dans un mo-

dèle *VARX* reste une variable de très faible influence. Les mémoires utiles de ces variables sont assez longues : 15h pour (T^{cor} et T^{above}) et 20h pour T^{ext}. Même sans T^{off} le bâtiment garde une grande inertie. La température de la pièce intérieure dépend des températures qu'il faisait 20 heures auparavant.

Remarque 15. *Si l'on adopte un point de vue à court terme, c'est-à-dire en ne projetant que sur les instants* $(t, t-1, t-2)$ *(indice* $S_{t,2}$*) ou* $(t, t-1)$ *(indice* $S_{t,1}$*),* T^{below} *est la plus influente. Ces indices se lisent dans les premiers points de la courbe. Dans le modèle* VARX *si l'on considère l'ensemble des variations de l'instant* 0 *à* t*,* T^{below} *n'est pas influente. Elle est supplantée par les fortes variations de* T^{ext} *et* T^{above}*.*

Si l'on considère un modèle de régression III.3 la variable la plus influente est la température du couloir. Les autres se trouvent à peu près au même niveau. Dans un modèle de régression, la sortie Y ne prend en considération que quelques instants passés des entrées. Il suggère un comportement à court terme. On peut en déduire, alors, que l'influence de T^{cor} est plus importante à court terme.

	T_t^{below}	T_t^{above}	T_t^{off}	T_t^{cor}	T_t^{ext}
S1-A	0.02	0.38	0.12	0.22	0.5
S2-A	0.05	0.55	-	0.28	0.73
S2-B	0.45	0.40	-	0.81	0.48

TABLE 5.2 – Tableau récapitulatif des indices de Sobol pour les différents modèles été

Remarque 16. *La somme des indices n'est pas inférieure ou égale à 1 car les variables sont dépendantes. Par contre chaque indice doit être inférieur ou égal à 1.*

5.2.2 Modèle hiver

Nous prenons dans ce modèle le chauffage comme sortie et cherchons à savoir quelle variable a le plus d'influence pour pouvoir gérer ou optimiser le flux de chauffage. Ce modèle prend en compte les températures des pièces chauffées. Le processus étant cyclo-stationnaire $S_{t,k}$ dépend de t de manière périodique pour toutes les mémoires k. Nous allons donc tracer pour chaque t fixé l'indice de sensibilité pour chaque variable. La périodicité implique cependant que pour une mémoire fixée k, l'indice est périodique : $S_{t+hP} = S_t, \quad \forall h \in \mathbb{Z}$ où P est la période qui dans notre cas est 24. On ne donnera alors que les indices pour $t = [\![0, 23]\!]$. L'indice converge vers une valeur, on ne donne que les valeurs finales et elles sont résumées dans le tableau 5.3 et dessinées figure III.4.

On constate premièrement qu'à chaque pas de temps l'indice converge bien vers une constante. On peut aussi remarquer que les variables influentes varient avec l'heure.
Pendant les heures de la nuit : "00h00 - 6h00" et "21h00 - 23h00", les variables les plus influentes sont T^{ext} et T^{above}. Pendant ces heures le chauffage (la sortie) fonctionne au ralenti. Ces variables ont les plus grandes variances et correspondent aux lieux les plus froids.

Lorsque le chauffage reprend à 8h00 et ce jusqu'à 12h00, les variables les plus influentes sont : T^{int} la pièce chauffée et T^{below} la pièce en dessous. Grâce au chauffage, la pièce de T^{int} doit atteindre une certaine température. Il est donc normal que cette pièce influence les variations du chauffage directement. On remarque que cette pièce est influente durant 4h, ce qui suggère que la pièce met un certain temps à se réchauffer.
A partir de 13h00 jusqu'à 18h00 la puissance de chauffage décroit légèrement : la variable ayant le plus d'impact est T^{cor} suivie de très près par T^{above}. Les effets des autres variables sont néanmoins très proches. T^{above} et T^{cor} sont les pièces directement en contact, froides et donc susceptibles d'engendrer des variations de chauffage.
Le chauffage est coupé à 18h00. A 19h00 la variable la plus influente est le nombre de personnes. Durant les heures de chauffage l'équivalent chaleur des étudiants N n'a pas d'impact sur les variations de la puissance de chauffage. Les apports ne sont soit pas suffisants, soit le chauffage est bloqué à son maximum. En effet l'air à chauffer est une fonction de l'air intérieur et de l'air prélevé à l'extérieur. Si il fait très froid, la puissance fournie sera très importante et peut atteindre son maximum. On peut, peut être voir là une mauvaise gestion du chauffage ou un mauvais choix du mode du chauffage. En journée, lors de la présence des étudiants, le chauffage devrait être moins important.

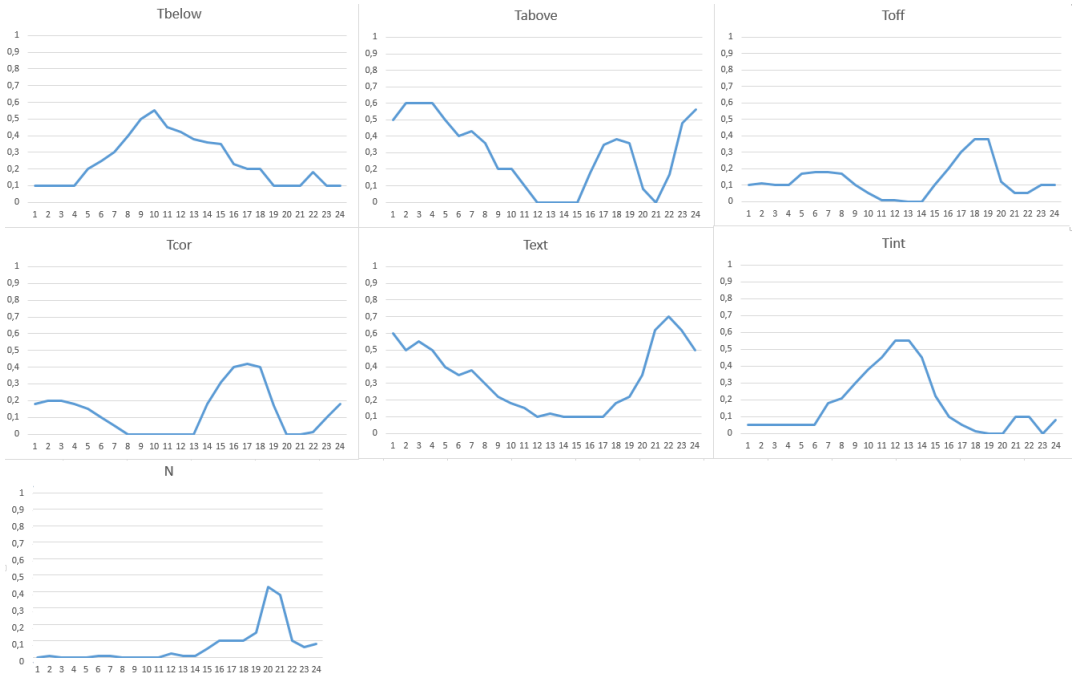

FIGURE III.4 – Indices de sensibilité pour chaque variable en fonction de l'heure estimés avec un échantillon de taille $N = 700$

Heure	Tbelow	Tabove	Toff	Tcor	Text	Tint	N
0	0,1	0,5	0,1	0,18	0,6	0,05	0
1	0,1	0,6	0,11	0,2	0,5	0,05	0,01
2	0,1	0,6	0,1	0,2	0,55	0,05	0
3	0,1	0,6	0,1	0,18	0,5	0,05	0
4	0,2	0,5	0,17	0,15	0,4	0,05	0
5	0,25	0,4	0,18	0,1	0,35	0,05	0,01
6	0,3	0,43	0,18	0,05	0,38	0,18	0,01
7	0,4	0,36	0,17	0	0,3	0,21	0
8	0,5	0,2	0,1	0	0,22	0,3	0
9	0,55	0,2	0,05	0	0,18	0,38	0
10	0,45	0,1	0,01	0	0,15	0,45	0
11	0,42	0	0,01	0	0,1	0,55	0,02
12	0,38	0	0	0	0,12	0,55	0,01
13	0,36	0	0	0,18	0,1	0,45	0,01
14	0,35	0	0,1	0,31	0,1	0,22	0,05
15	0,23	0,18	0,2	0,4	0,1	0,1	0,1
16	0,2	0,35	0,3	0,42	0,1	0,05	0,1
17	0,2	0,38	0,38	0,4	0,18	0,01	0,1
18	0,1	0,36	0,38	0,17	0,22	0	0,15
19	0,1	0,08	0,12	0	0,35	0	0,43
20	0,1	0	0,05	0	0,62	0,1	0,38
21	0,18	0,17	0,05	0,01	0,7	0,1	0,1
22	0,1	0,48	0,1	0,1	0,62	0	0,06
23	0,1	0,56	0,1	0,18	0,5	0,08	0,08

TABLE 5.3 – Indices de sensibilité pour chaque mémoire estimés avec un échantillon de taille $N = 700$

Chapitre 6

Conclusion

Un des enjeux de l'analyse de sensibilité en énergique des bâtiments est l'optimisation de l'efficacité énergétique notamment par une meilleure gestion. La plupart des études se sont déroulées dans un cadre statique et ne prennent pas en compte la dépendance des variables.

Différents modèles existent pour faire cette étude : modèle analytique (circuit électrique), modèle numérique (logiciel COMFIE notamment) qui réalise un bilan thermique et décompose le bâtiment en maille ou des modèles stochastiques, type séries chronologiques. Nous avons privilégiés ces derniers pour notre application.

L'application proposée est une plateforme expérimentale comportant une salle de classe. Le peu de données disponibles nous a poussé à étudier deux périodes de l'année : un mois durant l'été et un mois d'hiver. Cette application illustre comment on peut utiliser les indices de Sobol temporels à des fins de gestion et de l'importance de la notion pratique de mémoire utile dans les calculs de sensibilité. Par exemple cette étude a montré le comportement différent entre l'hiver et l'été du bâtiment. Le couloir (T^{cor}) joue un rôle essentiel. On a pu constater en été que cette pièce froide pourrait aider à refroidir la pièce et donc permettre un meilleur confort. Le temps de mise en action de cet effet peut prendre du temps et donc suggère la mise en place de ventilation par exemple naturelle pour accélérer les transferts thermiques.
Cet indice dynamique au cours du temps évolue et change. Particulièrement lors de la mise en place de protocoles, l'indice peut être utilisé pour détecter un défaut. Par exemple, en hiver, on a pu voir que la chaleur des utilisateurs n'est pas utilisée ou l'on ne voit pas son utilisation car le chauffage est bloqué à son maximum. Se dissipant rapidement ou du moins ayant un effet à très court terme sur les variations de la puissance de chauffage fournie, il aurait été plus utile de mettre une puissance de chauffage moins importante la journée pour tirer profit de celle des visiteurs. L'influence du nombre de personnes dans la pièce joue un rôle important. L'influence des apports internes tels que ceux liés à l'occupation est très importante, et doivent faire partie des variables prises en compte dans les systèmes de gestion énergétique [19]. L'extinction du chauffage peut, par exemple être mise en place automatiquement plusieurs minutes avant le déclenchement du thermostat, par anticipation d'une hausse de température. La forte variation des indices suivant l'heure en hiver montre que la gestion du chauffage doit tenir compte assez finement de cette procédure. On peut aussi penser à changer le type de chauffage utilisé. Celui ci dépend trop de la température qu'il fait à l'extérieur, souvent bloqué à son maximum il

ne permet pas de faire usage des apports d'autres types dans la pièce et donc de faire des économies.

Une des perspectives principales sera de mener une analyse couplée des paramètres des matrices représentant les caractéristiques physiques du bâtiment et ces variables dynamiques.

Quatrième partie

Conclusion et perspectives

Dans cette thèse, nous avons développé des méthodes d'analyse de sensibilité en vue de les appliquer à des problèmes liés à la conception et à la gestion de l'énergie dans les bâtiments. Dans ce domaine des variables d'entrée sont fortement corrélées et les méthodes existantes examinées dans les chapitres du début sont le plus souvent applicables sous la seule hypothèse d'indépendance.

Cela nous a amené du point de vue statistique à développer des techniques nouvelles pour tenir compte du caractère des entrées de type dynamique et corrélées entre elles à la fois dans le temps et à chaque instant. Nous avons placé notre travail dans l'optique de méthodes permettant de se ramener au cas d'entrées indépendantes. Notre préoccupation a été de modéliser les entrées de manière souple, aisément transposables à d'autres situations concrètes et permettant des simulations relativement aisées. Les formes des relations entrée-sortie ont peu d'importance, comme nous l'avons montré sur des cas d'école, sous la seule contrainte, évidemment non anodine, d'une simulation possible. Nous avons essayé aussi d'analyser plus en détail les problématiques liées à la sensibilité dans le cas du bâtiment. Nous avons relevé plusieurs problématiques, détaillées dans ce document et nous n'avons pu traiter en profondeur que certaines de ces problématiques parmi les plus importantes. L'obstacle principal dans notre travail a été les données d'observation dont nous avons pu disposer. La quantité insuffisante et la qualité de certaines de ces données ont impliqué de fortes difficultés dans les problèmes purement statistiques de choix de modèle et d'estimation des paramètres. Cela ne nous a pas permis d'utiliser les méthodes qui nous semblaient les plus performantes. Mais nos algorithmes, très efficaces, permettent de traiter sans modification des jeux de données beaucoup plus conséquents. Notons aussi les difficultés, celles-là intrinsèques, dues à de très fortes colinéarités entre variables d'entrée mais aussi entre celles-ci et la variable de sortie s'agissant de variables de températures internes. Nous avons pu régler ce problème de manière satisfaisante par l'utilisation simultanée de modèles « emboités ».

La première méthode développée, généralisation de la méthode Pick and Freeze, a reposé sur la définition des indices de Sobol adaptée au cadre dynamique. Afin de rendre compte du lien temporel entre les variables, nous avons choisi de considérer un indice dépendant, dans le cas non stationnaire (en particulier s'il existe des phénomènes saisonniers), de l'instant de calcul et de quantifier la variabilité de la sortie non pas seulement à la variabilité de l'entrée à l'instant t mais aussi à cette même variabilité provenant des instants précédents. Cette vision permet d'introduire la notion de mémoire utile pour le calcul de la sensibilité. Elle renvoie par exemple à des notions d'inertie thermique ou de vitesse de refroidissement/réchauffement. Si nous avions pu modéliser plus finement la situation et avions pu disposer de plus de données d'observation, des approches heuristiques nous montrent que la mémoire utile serait, sans doute de 15 heures dans la majorité des cas pour l'application Predis. Cette notion donne une indication sur l'attitude à adopter. On peut vouloir minimiser cette mémoire ou à l'inverse vouloir qu'elle soit importante. Par exemple en été, on voudrait minimiser cette mémoire utile afin d'accélérer les échanges thermique entre le couloir et la pièce que nous avons étudiée.

Dans le cadre d'entrées corrélées, les indices de Sobol ne sont plus associés à la décomposition d'Hoefding qui n'est généralisable que sous certaines conditions. Ces indices ne sont plus de somme égale à 1. L'indice associé à chaque variable ou chaque groupe de variables reste néanmoins inférieur à 1 et garde la même interprétation : plus cet indice est proche de 1, plus la variable est influente.

Facile à mettre en œuvre, la méthode d'estimation des indices de sensibilité développée ne s'appuie pas sur des hypothèses d'indépendance des entrées. Elle permet alors un large éventail d'applications et plus particulièrement dans le domaine appliqué où il est rare de posséder des données indépendantes. La seconde méthode que nous avons développée est une méthode d'estimation des indices de Sobol pour des entrées corrélées statiques a priori. Elle peut néanmoins être mise en œuvre pour des entrées dynamiques de courte mémoire mais les calculs sont alors lourds dès que le nombre d'entrées est grand ou les mémoires importantes. Cette méthode permet de transformer des variables dépendantes de loi quelconque en des variables indépendantes de loi uniforme. Il est alors facile de simuler de manière efficace un grand nombre de variables de loi uniforme et d'appliquer la méthode Pick and Freeze pour estimer ces indices. Elle est facile à mettre en œuvre si la loi de l'ensemble des variables est une loi donnée. Lorsqu'elle est inconnue et qu'il faut l'estimer nous avons réduit le problème à une suite de problèmes unidimensionnels assez bien connus en statistique non paramétrique. Comme dans beaucoup de problèmes sur la sensibilité, l'ordre dans lequel on traite des variables est important. L'ordre choisi pour calculer l'indice d'ordre 1 pour la variable X est un ordre devant débuter par X. Cette méthode est très bien adaptée au calcul des indices par blocs, en particulier de longueur importante. Nous avons commencé, et c'est une de nos perspectives, à étendre le travail fait sur les processus gaussiens classiques à des processus ayant la même structure de covariance mais de lois marginales très différentes des gaussiennes : lois bornées ou bimodales ou à queues lourdes ... Pour cela il faut construire des transformations types copules gaussiennes conservant ou adaptant les structures de covariance. Nous avons montré qu'au moins pour les indices d'ordre 1, ces transformations permettent de calculer l'indice Pick and Freeze en utilisant les méthodes développées pour les processus gaussiens. Nous voudrions étendre la famille des marginales que l'on peut considérer pour des processus VAR et comprendre les covariances atteignables pour ces méthodes, et pour finir étendre les calculs de sensibilité à des ordres supérieurs à 1.

Nous avons par la suite appliqué la première méthode à un problème du bâtiment. Nous avons fait le choix pour notre étude d'utiliser des modèles entrée-sortie linéaires de type séries temporelles (auto-régressions ou régressions linéaires) liés aux modèles d'état souvent utilisés dans les modélisations des bâtiments. Les entrées ont été modélisées par des processus Gaussiens multivariés du type $VAR(p)$ (vectorial autoregressive de mémoire p) transformés ensuite en $VAR(1)$ de dimension très supérieure. Nous avons choisi de détailler l'ordre 2 uniforme sur l'ensemble des variables. On aurait pu envisager des modélisations plus complexes (différents ordres pour les variables par exemple) mais le peu de données (30 jours de données), les fortes saisonnalités persistantes (saisonnalité de 24 heures) ou encore les changements de protocoles (par exemple la variable de chauffage) nous ont poussé à utiliser des modèles plus simples. Nous avons pu montrer différentes situations en analysant l'ordre des variables suivant les sensibilités. L'effet de la saisonnalité des variances et l'effet de très grandes corrélations entre variables compliquent la discussion, en particulier pour déterminer la taille des mémoires efficaces.

L'application de cette méthode aux problèmes du bâtiment permet de prendre en considération aussi bien l'aspect dynamique des entrées (températures de pièces, chauffages, présence de personnes, inertie des échanges thermiques) que celle de la sortie.

Il est évident que cette méthode appliquée à un bâtiment existant peut aider à une meilleure gestion de l'énergie. Il peut être intéressant de se rendre compte que le nombre de personnes

dans la pièce participe au chauffage. Cette variable peut être utile pour la gestion de celui ci. On peut par exemple anticiper son action : on peut imaginer de le couper lorsque le nombre de personnes dans la pièce est important. Cette méthode pourrait être utile en conception à partir de la mise en œuvre de scénarios. Le bâtiment est représenté en tenant compte de la connaissance des matériaux utilisés et de leurs qualités. Les températures d'entrées des pièces peuvent être modélisées à partir de connaissances passées, en prenant un exemple de scénario, par exemple de vouloir que le chauffage d'une pièce se fasse à partir de celle adjacente. Si l'analyse de sensibilité révèle que celle-ci n'est pas influente alors il faudra peut-être changer la conception ou la mise en œuvre de la gestion.

Une des perspectives futures que nous avons amorcée mais non achevée faute de données suffisantes, est d'effectuer une analyse des paramètres physiques utilisés pour la conception du bâtiment, paramètres eux-mêmes représentés par des lois de probabilité choisies de manière gaussiennes multivariées à partir des observations. Cette approche purement statistique permet de quantifier les incertitudes des paramètres liées par exemple à certains matériaux autour d'une valeur nominale connue. Nos méthodes permettent de prendre en considération la corrélation des variables et celle des paramètres, résultat intéressant en soi. Il est donc possible de coupler l'analyse de sensibilité des entrées dynamiques aux entrées statiques (paramètres de conception) et ainsi de se rendre compte de l'impact à chaque instant du lien entre ces différentes variables. Le travail numérique reste à faire dès que l'on pourra disposer d'un jeu de données plus important. Un modèle envisagé pour traiter ce type de problème était d'utiliser un modèle d'état du bâtiment où les paramètres des matrices, représentant les caractéristiques de conception, étaient estimés par filtre de Kalman. A partir d'un bootstrap sur les résidus on estime la loi des paramètres afin d'effectuer une analyse de sensibilité. Malheureusement les fortes saisonnalités et les réglages difficiles n'ont pour l'instant pas permis d'aboutir à des résultats satisfaisants. Le réglage du filtrage s'est avéré très difficile. Peut-être est-ce dû au choix du modèle entrée-sortie et à ses très fortes colinéarités.

Une étude envisageable sera aussi de regarder comment se comporte le bâtiment dans des températures extrêmes. A cet effet il faudrait disposer de données assez conséquentes associées aux pointes ou aux vagues de chaleur et de froid. Nous pensons avoir développé deux méthodes qui jointes, pourraient permettre une analyse de sensibilité pour le bâtiment qui soit globale, c'est-à-dire simultanée sur les paramètres physiques liés aux matériaux et sur les processus stochastiques d'entrée que sont en particulier les différentes températures d'intérêt, le chauffage et la présence de visiteurs. Quant aux résultats théoriques nouveaux ils montrent la capacité de la méthode Pick and Freeze à traiter simultanément la dimension temporelle avec des entrées dépendantes, par exemple pour le cas de processus AR non gaussiens ayant des propriétés qualitatives spécifiques sur les supports, la bi-modalité, l'asymétrie, la lourdeurs des queues. Nous souhaitons développer cette approche.

Appendices

Annexe A

Modèles d'entrée été

Cette partie présentent les valeurs des coefficients obtenue pour les différents modèles d'entrées utilisé en partie : 4.3.2.

A.0.1 Modèle E-A multivarié

Le modèle est de la forme :

$$\boldsymbol{U}_t = \sum_{k=1}^{p} A_k \boldsymbol{U}_{t-k} + \varepsilon_t \quad \text{avec } \varepsilon \sim \mathcal{N}(0, \Gamma)$$

	T_t^{below}	T_t^{above}	T_t^{off}	T_t^{cor}	T_t^{ext}
T_{t-1}^{below}	0.88	-0.23	0.21	0.12	0.20
T_{t-2}^{below}	0.04	0.03	-0.16	-0.03	-0.45
T_{t-1}^{above}	0.01	1.21	0.03	0.06	0.61
T_{t-2}^{above}	-0.01	-0.48	-0.02	-0.04	-0.60
T_{t-1}^{off}	0.06	0.01	1.24	0.03	-0.09
T_{t-2}^{off}	-0.07	-0.04	-0.3	-0.05	0.25
T_{t-1}^{cor}	0.06	0.49	0.06	1	0.71
T_{t-2}^{cor}	-0.04	-0.08	-0.09	-0.15	-0.31
T_{t-1}^{ext}	0	0.3	0.01	0.01	0.85
T_{t-2}^{ext}	0.01	-0.04	-0.1	-0.01	-0.05

TABLE A.1 – Coefficients du **modèle E-A**

A.0.2 Modèle E-B : Modèle multivarié sans T^{off}

Dans ce modèle, nous suggérons de réunir le bureau et la pièce d'intérêt en une seule pièce d'un point de vue thermique.

T^{below}	0.01	0.01	0.004	0.006	0.01
T^{above}	0.01	0.62	0.02	0.03	0.46
T^{off}	0.004	0.02	0.03	0.002	0.04
T^{cor}	0.006	0.03	0.002	0.02	0.04
T^{ext}	0.01	0.46	0.04	0.04	1.34

TABLE A.2 – Matrice de covariance des bruits modèle E-A

T^{below}	1	0.12	0.24	0.33	0.07
T^{above}	0.12	1	0.20	0.29	0.51
T^{off}	0.24	0.20	1	0.07	0.23
T^{cor}	0.33	0.29	0.07	1	0.21
T^{ext}	0.07	0.51	0.23	0.21	1

TABLE A.3 – Matrice de corrélation des bruits modèle E-A

Le nouveau $\boldsymbol{U}_t$, notée U_t^1 est :

$$U_t^1 = (T_t^{\text{above}}, T_t^{\text{below}}, T_t^{\text{cor}}, T^{\text{ext}})$$

On peut comme précédemment considérer un modèle multivariée *VAR* .

	T^{below}	T^{above}	T^{cor}	T^{ext}
T_{t-1}^{below}	0.90	-0.23	0.12	0.19
T_{t-2}^{below}	0.01	0.01	-0.07	-0.38
T_{t-1}^{above}	0.02	1.21	0.07	0.61
T_{t-2}^{above}	-0.01	-0.48	-0.05	-0.59
T_{t-1}^{cor}	0.07	0.50	1.02	0.68
T_{t-2}^{cor}	-0.03	-0.08	-0.5	-0.31
T_{t-1}^{ext}	0.001	0.14	0.01	0.85
T_{t-2}^{ext}	0.002	-0.04	-0.01	-0.04

TABLE A.4 – Coefficients du **modèle E-B**

T^{below}	0.01	0.01	0.01	0.01
T^{above}	0.01	0.62	0.04	0.46
T^{cor}	0.01	0.04	0.03	0.04
T^{ext}	0.01	0.46	0.04	1.35

TABLE A.5 – Matrice de covariance des bruits modèle**E-B**

T^{below}	1	0.12	0.34	0.07
T^{above}	0.12	1	0.29	0.51
T^{cor}	0.34	0.29	1	0.20
T^{ext}	0.07	0.51	0.2 0	1

TABLE A.6 – Matrice de corrélation des bruits modèle **E-B**

Annexe B

Filtre de Kalman

B.1 Filtre de Kalman

On peut comprendre le mot filtre de deux façons. La première consiste à enlever le bruit d'une quantité mesurée. La seconde consiste à retrouver à l'instant t de l'information sur le système $s(.)$ à partir des mesures disponibles jusqu'à l'instant t.

Il faut distinguer encore deux types d'estimations dans le traitement de l'information : la prédiction et le lissage.

Le lissage consiste à rechercher la meilleure approximation du signal sachant les observations passées, présentes et futures.

La prédiction consiste à prévoir la valeur du signal à un instant $(t+\lambda)$, $\lambda > 0$, à partir des mesures disponibles jusqu'à l'instant t.

Le filtre de Kalman permet d'estimer ou de prédire le vecteur d'état α_t. A chaque instant t, on va chercher à estimer les variables d'état conditionnellement aux variables observées jusqu'à la date $t : Y_t = (y_1, \ldots, y_t)$.

B.1.1 Prérequis

Soit (x, y) un vecteur gaussien tel que

$$\mathbf{E}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \tag{IV.1}$$

et

$$\mathbf{Var}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{yy} \end{pmatrix} \tag{IV.2}$$

Lemme 3. *La distribution conditionnelle de x sachant y est une loi normale de moyenne :*

$$\mathbf{E}(x|y) = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) \tag{IV.3}$$

et de variance :

$$\mathbf{Var}(x|y) = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1T}\Sigma_{xy} \tag{IV.4}$$

On considère maintenant que y est connu (par exemple observé) et x inconnu.

Un estimateur de x peut s'écrire :

$$\hat{x} = \mathbf{E}(x|y) \tag{IV.5}$$

Lemme 4. *Soit (x, y) un vecteur Normal, alors l'estimateur de x, $\hat{x} = \mathbf{E}(x|y)$ est un estimateur linéaire sans biais et de variance minimale connaissant y et la variance de son erreur est donnée par l'équation : (IV.4).*

B.1.2 Filtre de Kalman

Le filtre de Kalman calcule de manière récursive un estimateur du vecteur d'état $\hat{\alpha}_t$, linéaire sans biais et de variance minimale.

Dès lors que l'on a une nouvelle observation y_t, on peut calculer α_{t+1}.

Supposons que l'on ai déjà observé les t premières valeurs de y. On va les noter $Y_t = (y_1, \ldots, y_t)$

On peut définir différents problèmes :

- Estimer α_t conditionnellement à Y_{t-1} définit un problème de prédiction
- Estimer α_t conditionnellement à Y_t définit un problème de filtrage
- Estimer α_t conditionnellement à $Y_n \quad n > t$ définit un problème de lissage

Le filtre de Kalman va opérer en deux temps. Le premier est une phase de prédiction :

$$a_{t+1} \quad = \quad \mathbf{E}(\alpha_{t+1}|Y_t) \;\; et \;\; P_{t+1} = \mathbf{Var}(\alpha_{t+1}|Y_t) \tag{IV.6}$$

puis dans un second temps on procède à une phase de filtrage :

$$a_{t|t} \quad = \quad \mathbf{E}(\alpha_t|Y_t) \;\; et \;\; P_{t|t} = \mathbf{Var}(\alpha_t|Y_t) \tag{IV.7}$$

a_{t+1} est le meilleur estimateur linéaire de α_{t+1}.

On définit aussi v_t l'innovation ou l'erreur de prédiction de y_t connaissant Y_{t-1} :

$$v_t = y_t - \mathbf{E}(y_t|Y_{t-1}) = y_t - \mathbf{E}(Z_t\alpha_t + \epsilon_t) = y_t - Z_t a_t \tag{IV.8}$$

Quand v_t et Y_{t-1} sont fixés alors Y_t l'est. On peut alors réécrire :

$$a_{t|t} \quad = \quad \mathbf{E}(\alpha_t|Y_t) = \mathbf{E}(\alpha_t|Y_{t-1}, v_t) \tag{IV.9}$$

$$a_{t+1} \quad = \quad \mathbf{E}(\alpha_{t+1}|Y_t) = \mathbf{E}(\alpha_{t+1}|Y_{t-1}, v_t) \tag{IV.10}$$

On se place à un instant t. On a prédit la valeur a_t et P_t :

phase de filtrage

La distribution conditionnelle jointe de α_t et v_t connaissant Y_{t-1} est d'après le lemme 3

$$a_{t|t} = \mathbf{E}(\alpha_t|Y_{t-1}) + \mathbf{Cov}(\alpha_t, v_t)\mathbf{Var}^{-1}(v_t|Y_{t-1})v_t \tag{IV.11}$$

On peut calculer la $\mathbf{Cov}(\alpha_t, v_t)$ conditionnellement à Y_{t-1} :

$$\begin{aligned}
\mathbf{Cov}(\alpha_t, v_t) &= \mathbf{E}(\alpha_t v_t'|Y_{t-1}) \\
&= \mathbf{E}(\alpha_t(Z_t\alpha_t + \epsilon_t - Z_t a_t)'|Y_{t-1}) \\
&= \mathbf{E}(\alpha_t(\alpha_t - a_t)'Z_t'|Y_{t-1}) \\
&= \mathbf{Var}(\alpha_t|Y_{t-1}) \\
&= P_t Z_t'
\end{aligned}$$

On note F_t la variance de v_t conditionnellement à Y_{t-1}. D'après le lemme 3

$$\begin{aligned}
F_t &= \mathbf{Var}(v_t|Y_{t-1}) \\
&= \mathbf{Var}(Z_t\alpha_t + \epsilon_t - a_t Z_t|Y_{t-1}) \\
&= Z_t P_t Z_t' + H_t
\end{aligned}$$

Alors on peut réécrire (IV.11) en supposant que F_t est inversible par :

$$a_{t|t} = a_t + P_t^T F_t^{-1} v_t \tag{IV.12}$$

Remarque 17. *On a supposé que F_t est inversible. Cette hypothèse est vraie dans la plupart des modèles. Dans certains cas on peut assouplir cette hypothèse.*

Calcul de la variance conditionnelle :

$$\begin{aligned}
P_{t|t} &= \mathbf{Var}(\alpha_t|Y_t) \\
&= \mathbf{Var}(\alpha_t|Y_{t-1}, v_t) \\
&= \mathbf{Var}(\alpha_t|Y_{t-1}) + \mathbf{Cov}(\alpha_t, v_t)\mathbf{Var}^{-1}(v_t|Y_{t-1})^T\mathbf{Cov}(\alpha_t, v_t) \\
&= P_t + P_t^T Z_t F_t^{-1} Z_t P_t
\end{aligned} \tag{IV.13}$$

Phase de prédiction :

$$\begin{aligned}
a_{t+1} &= \mathbf{E}(\alpha_{t+1}|Y_t) \\
&= \mathbf{E}(T_t\alpha_t + R_t\eta_t|Y_t) \\
&= T_t\mathbf{E}(\alpha_t|Y_t) \\
&= T_t a_{t|t}
\end{aligned} \tag{IV.14}$$

Remarque 18. *Il suffit de programmer $a_{t|t}$ pour obtenir a_{t+1}*

Grâce à l'équation (IV.12) on obtient alors

$$a_{t+1} = T_t a_t + T_t P_t Z_t' F_t^{-1} v_t \tag{IV.15}$$

On note $K_t = T_t P_t Z_t' F_t^{-1}$.

La matrice K_t est appelée le gain de Kalman.

On calcule la variance conditionnelle :

$$\begin{aligned} P_{t+1} &= \mathbf{Var}(\alpha_{t+1}|Y_t) \\ &= \mathbf{Var}(T_t\alpha_t + R_t\eta_t|Y_t) \\ &= T_t\mathbf{Var}(\alpha_t|Y_t)'T_t + R_tQ_tR_t' \\ &= T_tP_{t|t}T_t' + R_tQ_tR_t' \\ &= T_tP_t(T_t - K_tZ_t)' + R_tQ_tR_t' \end{aligned} \tag{IV.16}$$

B.1.3 Initialisation du filtre de Kalman

Jusqu'à présent on a supposé que :

$$\alpha_1 \sim N(a_1, P_1) \tag{IV.17}$$

avec a_1 et P_1 supposés connus.

Dans la plupart des applications, certains ou tous les éléments de a_1 et P_1 sont inconnus. Nous allons présenter dans la suite une méthode dite d'initialisation.

Nous considérons dans la suite que α_1 possède certains éléments connus (la distribution jointe est connue) alors que certains sont totalement inconnus.

Un des modèles généraux pouvant être considéré pour le paramètre d'état initial α_1 est :

$$\alpha_1 = a + A\delta + R_0\eta_0, \quad \eta_0 \sim N(0, Q_0) \tag{IV.18}$$

où :

- a est connu. Dans la plupart des cas a est considéré comme nul
- δ est une quantité inconnue
- A et R_0 sont des matrices choisies contenant des colonnes de la matrice identité I
- Q_0 est supposée définie positive et connue

Cette représentation a pour but de séparer α_1 en :

- une partie constante : a
- une partie non stationnaire : $A\delta$
- une partie sationnaire : $R_0\eta_0$

Il nous reste à déterminer δ.

δ peut être traité comme un vecteur fixe aux paramètres inconnus ou comme un vecteur de variables aléatoires suivant une loi normale de variance infinie.

Dans le cas où δ est un vecteur fixe, De Jong ([27]) utilise une méthode de maximum de vraisemblance afin d'estimer δ. Nombre d'analystes préfèrent cette méthode pour le simple fait qu'il n'existe pas de variable physique équivalente possédant de variance infinie.

On utilise la méthode diffuse dans le cas non stationnaire. Dans le cas stationnaire on a $P_{inf} = 0$.

Méthode d'initialisation On considère δ comme un vecteur aléatoire :

$$\delta \sim N(0, \kappa I) \quad \kappa \to \infty \tag{IV.19}$$

Le filtre de Kalman a pour condition initiale :

$$\begin{array}{rcccl} a_1 & = & \mathbf{E}(\alpha_1) & = & a \\ P_1 & = & \mathbf{Var}(\alpha_1) & = & \kappa AA' + R_0 Q_0 R_0' = \kappa P_\infty + P_* \quad \kappa \to \infty \end{array} \tag{IV.20}$$

Pour calculer P_1 on peut par exemple remplacer κ par un nombre très grand et utiliser le filtre de Kalman standard présenté plus haut (voir Harvey and Phillips ([56])). Cette technique est utilisée dans un premier temps pour un travail exploratoire. Elle n'est cependant pas recommandée dans le cas général car elle conduit à une trop grande erreur d'approximation et à des instabilités numériques.

On peut montrer de manière identique que P_t et F_t peuvent se décomposer sous la forme :

$$P_t = \kappa P_{\infty,t} + P_{*,t} \tag{IV.21}$$

$$F_t = \kappa F_{\infty,t} + F_{*,t} \tag{IV.22}$$

On utilise cette forme pour construire le filtre de Kalman.

On peut montrer qu'à partir d'un certain rang d, pour $t > d \quad P_{\infty,t} = 0$ et donc on retombe sur le filtre de Kalman "classique" à partir de $t = d + 1$.

B.2 Estimation des paramètres par maximum de vraisemblance

Jusqu'à présent nous avons développé des méthodes afin d'estimer les paramètres qui sont placés à l'intérieur du vecteur d'état.

En réalité, les modèles réels dépendent d'autres paramètres qui ont besoin d'être estimés. Ces paramètres peuvent être contenus dans les matrices Z_t ou T_t, ou correspondre aux variances Q_t ou H_t (II.24).

Lors de l'analyse classique, ces paramètres sont supposés fixes mais inconnus alors que d'un point de vue Bayésien ils sont supposés être des variables aléatoires.

Dans le cas du modèle linéaire Gaussien on peut montrer qu'une méthode de maximum de vraisemblance peut être appliquée au filtre de Kalman, même si l'état initial est diffus ou partiellement diffus.

Un outil pratique pour maximiser la vraisemblance est l'algorithme EM [85], particulièrement lors d'une première étape.

B.2.1 Calcul de la vraisemblance

Vraisemblance lorsque la condition initiale est connue
On suppose que l'état initial est connu, c'est-à-dire que :

$$\alpha_1 \sim N(a_1, P_1)$$

avec a_1 et P_1 connu.

On considère ψ le vecteur des paramètres inconnus.

Soit un échantillon $(y_1, \ldots, y_n)$ de n variables indépendantes alors par définition la vraisemblance est donnée par :

$$l(\psi, y_1, \ldots, y_n) = \mathbf{P}(Y = y_1)\Pi_{i=1}^{n}\mathbf{P}(y_i|Y_{i-1}) \tag{IV.23}$$

Avec $Y_{i-1} = (y_1, \ldots, y_{i-1})$.

On prend par convention : $\mathbf{P}(y_1) = \mathbf{P}(y_1|Y_0)$. On peut réécrire cette vraisemblance avec les densités de probabilité :

$$l(\psi, y_1, \ldots, y_n) \quad = \quad \Pi_{i=1}^{n} f(y_i|Y_{i-1}) \tag{IV.24}$$

On a supposé précédemment que ϵ_t , η_t , α_1 suivaient des lois Normales. Donc les densités de probabilité sont des lois Normales de moyenne et de variance :

$$\begin{cases} \mathbf{E}(y_i|Y_{i-1}) &= a_t Z_t \\ \mathbf{Var}(y_i|Y_{i-1}) &= F_t \end{cases}$$

Donc $f(y_i|Y_{i-1}) \sim N(a_t Z_t, F_t)$

On en déduit alors une forme simplifiée de la log-Vraisemblance :

$$\log l(\psi, Y_n) = Constante - \frac{1}{2}\Sigma_{i=1}^{n}(log(|F_t| + v_t' F_t^{-1} v_t) \tag{IV.25}$$

On remarque alors que la log-Vraisemblance est calculable directement par filtre de Kalman. En effet les quantités v_t et F_t sont calculées grâce à celui ci.

Pour calculer le maximum on va devoir résoudre l'équation $\dfrac{\partial \log l(\psi, Y_n)}{\partial \psi} = 0$

B.3 Algorithme EM

L'algorithme EM (Expectation-Maximisation) [85] est une procédure itérative qui permet de calculer un estimateur du maximum de vraisemblance lorsque seulement une partie des données est disponible.

Dans sa forme classique le vecteur des données "complètes" W est constitué du vecteur des données observées Y et celles non observées α.

Une phase $E - step$: Expectation et $M - step$: Maximisation sont mises en œuvre.

— On initialise le vecteur des paramètres inconnus : ψ.

— $E - step$: On applique le filtre de Kalman afin d'obtenir α_t à partir des y_t et des matrices précédemment définies grâce aux paramètre $\psi^{(i)}$.
On calcule l'espérance de la log-vraisemblance en tenant compte des dernières variables observées.

$$Q(\psi|\psi^{(i)}) = \mathbf{E}_{\psi^{(i)}}(l(\psi; Y, \alpha)|Y, \psi^{(i)}) \tag{IV.26}$$

— $M - step$: On estime le maximum de vraisemblance des paramètres en maximisant la vraisemblance trouvée à l'étape E.

$$\psi^{(i+1)} = \underset{\psi}{\mathbf{argmax}}\, Q(\psi|\psi^{(i)}) \tag{IV.27}$$

A cette étape on utilise une méthode de type Newton ou quasi Newton (pour accélérer la convergence) afin de trouver les zéros du gradient.
On met à jour les matrices $Z_t,\ T_t,\ R_t\ , H_t,\ Q_t$ en calculant le maximum de la vraisemblance.

— On répète ces deux phases jusqu'à la convergence de l'algorithme.

On pourrait se contenter de l'étape M, cependant on n'est pas assuré de trouver le maximum de la vraisemblance.

La force de cet algorithme est sa monotonie. Après chaque itération la vraisemblance augmente. C'est-à-dire :

$$l(\psi^{(i)+1}) \geq l(\psi^{(i)})$$

On est alors assuré de la convergence dès lors que la vraisemblance est majorée.

L'algorithme EM est stable numériquement et est généralement facile à programmer.

Dempster, Laird, and Rubin (1977) [31] montrent que cet algorithme converge vers un maximum local. On comprend alors aisément que le maximum trouvé dépend de l'initialisation de l'algorithme.

On peut aussi remarquer qu'il converge assez rapidement dans les premières itérations. Lorsqu'il s'approche du maximum celui ci devient beaucoup plus lent. C'est pour cela qu'à partir d'un certain rang on utilise une méthode de maximisation classique afin de trouver un meilleur maximum. Une discussion est donnée dans Watson et Engle [121].

Cinquième partie

Bibliographie

Bibliographie

[1] Paul L Anderson and Mark M Meerschaert. Parameter estimation for periodically stationary time series. *Journal of Time Series Analysis*, 26(4) :489–518, 2005.

[2] Anestis Antoniadis, Jacques Berruyer, and René Carmona. *Régression non linéaire et applications*. Economica, 1992.

[3] Jean-Marc Azaïs and Jean-Marc Bardet. *Le modèle linéaire par l'exemple-2e éd. : Régression, analyse de la variance et plans d'expérience illustrés avec R et SAS*. Dunod, 2012.

[4] Robert Azencott and Didier Dacunha-Castelle. *Séries d'observations irrégulières : modélisation et prévision*. Elsevier Masson, 1984.

[5] Robert Azencott and Didier Dacunha-Castelle. *Series of irregular observations : forecasting and model building*, volume 2. Springer Science & Business Media, 2012.

[6] Ilaria Ballarini and Vincenzo Corrado. Analysis of the building energy balance to investigate the effect of thermal insulation in summer conditions. *Energy and Buildings*, 52 :168–180, 2012.

[7] Jean-Marc Bardet and Jean-Marc Azaïs. *Le modèle linéaire par l'exemple : Régression, Analyse de la variance et Plans d'expŕience illustrés avec R, SAS et Splus*. Dunod, April 2006.

[8] Bahar Biller and Barry L Nelson. Modeling and generating multivariate time-series input processes using a vector autoregressive technique. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13(3) :211–237, 2003.

[9] G. Blatman and B. Sudret. Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. *Reliability Engineering and System Safety*, 95 :1216–1229, 2010.

[10] Geraud BLATMAN. *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis*. PhD thesis, Universitée Blaise Pascal, 2009.

[11] Hilde Breesch and Arnold Janssens. Performance evaluation of passive cooling in office buildings based on uncertainty and sensitivity analysis. *Solar energy*, 84(8) :1453–1467, 2010.

[12] Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. Springer Science & Business Media, 2006.

[13] Peter J. Brockwell and Richard A. Davis. *Time Series : Theory and Methods*. Springer, 2nd ed. 1991. 2nd printing 2009 edition, April 2009.

[14] G.T. Buzzard and D. Xiu. Variance-based global sensitivity analysis via sparse-grid interpolation and cubature. *Communications in Computational Physics*, 9 :542–567, 2011.

[15] Marne C Cario and Barry L Nelson. Autoregressive to anything : Time-series input processes for simulation. *Operations Research Letters*, 19(2) :51–58, 1996.

[16] Gaëlle Chastaing. *Indices de Sobol généralisés pour variables dépendantes*. PhD thesis, Université de Grenoble, 2013.

[17] Gaëlle Chastaing, Fabrice Gamboa, Clémentine Prieur, et al. Generalized hoeffding-sobol decomposition for dependent variables-application to sensitivity analysis. *Electronic Journal of Statistics*, 6 :2420–2448, 2012.

[18] Gaëlle Chastaing and Loic Le Gratiet. Anova decomposition of conditional gaussian processes for sensitivity analysis with dependent inputs. *Journal of Statistical Computation and Simulation*, 85(11) :2164–2186, 2015.

[19] Hervé Chenailler. *L'efficacité d'usage énergétique : pour une meilleure gestion de l'énergie électrique intégrant les occupants dans les bâtiments*. PhD thesis, Université de Grenoble, 2012.

[20] Jean-Paul Chiles and Pierre Delfiner. *Geostatistics : modeling spatial uncertainty*, volume 497. John Wiley & Sons, 2009.

[21] T. Crestaux and O. Le Maître J.-M. Martinez. Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering and System Safety*, 94 :161–1172, 2009.

[22] R.I. Cukier, C.M. Fortuin, K.E. Shuler, A.G. Petschek, and J.H. Schaibly. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. i. theory. *J. Chemical Physics*, 59 :3873–3878, 1973.

[23] R.I. Cukier, R.I. Levine, and K.E. Shuler. Nonlinear sensitivity analysis of multiparameter model systems. *J. Computational Physics*, 26 :1–42, 1978.

[24] R.I. Cukier, J.H. Schaibly, and K.E. Shuler. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. iii. analysis of the approximations. *J. Chemical Physics*, 63 :1140–1149, 1975.

[25] Sebastien Da Veiga, Francois Wahl, and Fabrice Gamboa. Local polynomial estimation for sensitivity analysis on models with correlated inputs. *Technometrics*, 51(4) :452–463, 2009.

[26] Didier Dacunha-Castelle and Marie Duflo. *Probabilités et statistiques : problèmes à temps fixe*, volume 1. Masson, 1982.

[27] Piet De Jong. The likelihood for a state space model. *Biometrika*, 75(1) :165–169, 1988.

[28] P De Wilde and W Tian. Preliminary application of a methodology for risk assessment of thermal failures in buildings subject to climate change. In *Building Simulation*, pages 2077–2084, 2009.

[29] Sten De Wit and Godfried Augenbroe. Analysis of uncertainty in building design evaluations and its implications. *Energy and Buildings*, 34(9) :951–958, 2002.

[30] Christine Demanuele, Tamsin Tweddell, and Michael Davies. Bridging the gap between predicted and actual energy performance in schools. In *World renewable energy congress XI*, pages 25–30, 2010.

[31] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[32] Fernando Domínguez-Muñoz, José M Cejudo-López, and Antonio Carrillo-Andrés. Uncertainty in peak cooling load calculations. *Energy and Buildings*, 42(7) :1010–1018, 2010.

[33] Virginie Dordonnat, Siem Jan Koopman, and Marius Ooms. Dynamic factors in periodic time-varying regressions with an application to hourly electricity load modelling. *Computational Statistics & Data Analysis*, 56(11) :3134–3152, 2012.

[34] Francois xavier LE Dumet and Olivier Talagrand. Variational algorithms for analysis and assimilation of meteorological observations : theoretical aspects. *Tellus A*, 38(2) :97–110, 1986.

[35] James Durbin and Siem Jan Koopman. *Time series analysis by state space methods*. Number 38. Oxford University Press, 2012.

[36] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, 1(1) :54–75, 1986.

[37] B. Efron and R. Tibshirani. *An introduction to bootstrap*. Chapman & Hall, 1 edition, 1993.

[38] Bryan Eisenhower, Zheng O'Neill, Vladimir A Fonoberov, and Igor Mezić. Uncertainty and sensitivity decomposition of building energy models. *Journal of Building Performance Simulation*, 5(3) :171–184, 2012.

[39] P. Enciu, F. Wurtz, L. Gerbaud, and B. Delinchant. Automatic differentiation for electromagnetic models used in optimization. *COMPEL : Int J for Computation and Maths. in Electrical and Electronic Eng.*, 28(5) :1313–1326, 2009.

[40] Jianqing Fan, I Gijbels, Tien-Chung Hu, and Li-Shan Huang. *An asymptotic study of variable bandwidth selection for local polynomial regression with application to density estimation*. Department of Statistics [University of North Carolina at Chapel Hill], 1993.

[41] Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications : monographs on statistics and applied probability 66*, volume 66. CRC Press, 1996.

[42] Jianqing Fan and Qiwei Yao. Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3) :645–660, 1998.

[43] Jianqing Fan, Qiwei Yao, and Howell Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1) :189–206, 1996.

[44] Lukáš Ferkl and Jan Širokỳ. Ceiling radiant cooling : Comparison of armax and subspace identification modelling methods. *Building and Environment*, 45(1) :205–212, 2010.

[45] Steven K Firth, Kevin J Lomas, and AJ Wright. Targeting household energy-efficiency measures using sensitivity analysis. *Building Research & Information*, 38(1) :25–41, 2010.

[46] Gilles Fraisse, Christelle Viardot, Olivier Lafabrie, and Gilbert Achard. Development of a simplified and accurate building model based on electrical analogy. *Energy and buildings*, 34(10) :1017–1031, 2002.

[47] Fabrice Gamboa, Alexandre Janon, Thierry Klein, and Agnès Lagnoux. Sensitivity indices for multivariate outputs. 2013.

[48] Fabrice Gamboa, Alexandre Janon, Thierry Klein, Agnes Lagnoux-Renaudie, and Clémentine Prieur. Statistical inference for sobol pick freeze monte carlo method. *arXiv preprint arXiv :1303.6447*, 2013.

[49] Soumyadip Ghosh and Shane G Henderson. Behavior of the norta method for correlated random vector generation as the dimension increases. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 13(3) :276–294, 2003.

[50] Jeanne Goffart. *Impact de la variabilité des données météorologiques sur une maison basse consommation. Application des analyses de sensibilité pour les entrées temporelles.* PhD thesis, Université de Grenoble, 2013.

[51] MM Gouda, S Danaher, and CP Underwood. Building thermal model reduction using nonlinear constrained optimization. *Building and Environment*, 37(12) :1255–1265, 2002.

[52] Mathilde Grandjacques, Alexandre Janon, Benoit Delinchant, and Olivier Adrot. Pick-freeze estimation of projection on the past sensitivity indices for models with dependent causal processes inputs. *arXiv preprint arXiv :1403.5539*, 2014.

[53] Peter Hall, Rodney CL Wolff, and Qiwei Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445) :154–163, 1999.

[54] Edward James Hannan. Time series analysis. 1960.

[55] Edward James Hannan. *Multiple time series*, volume 38. John Wiley & Sons, 2009.

[56] Andrew C Harvey and Gary DA Phillips. Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66(1) :49–58, 1979.

[57] Trevor Hastie, Robert Tibshirani, Jerome Friedman, T Hastie, J Friedman, and R Tibshirani. *The elements of statistical learning*, volume 2. Springer, 2009.

[58] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

[59] Shane G Henderson and Barry L Nelson. *Handbooks in Operations Research and Management Science : Simulation : Simulation*, volume 13. Elsevier, 2006.

[60] Thi Thu Huong Hoang. *Modélisation de séries chronologiques non stationnaires, non linéaires : application à la définition des tendances sur la moyenne, la variabilité et les extrêmes de la température de l'air en Europe.* PhD thesis, Paris 11, 2010.

[61] Giles Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3), 2007.

[62] Christina J Hopfe and Jan LM Hensen. Uncertainty analysis in building performance simulation for design support. *Energy and Buildings*, 43(10) :2798–2805, 2011.

[63] S. huang, S. Mahadevan, and R. Rebba. Collocation-based stochastic finite element analysis for random field problems. *Probabilistic Engineering Mechanics*, 22 :194–205, 2007.

[64] G Hudson and CP Underwood. A simple building modelling procedure for matlab/simulink. In *Proceedings of the International Building Performance and Simulation Conference, Kyoto Japan*, volume 2, pages 777–783, 1999.

[65] Janelle S Hygh, Joseph F DeCarolis, David B Hill, and S Ranji Ranjithan. Multivariate regression as an energy assessment tool in early building design. *Building and Environment*, 57 :165–175, 2012.

[66] Bertrand Iooss. Revue sur l'analyse de sensibilité globale de modèles numériques. *Journal de la Société Française de Statistique*, 152(1) :3–25, 2011.

[67] Bertrand Iooss, Loïc Boussouf, Vincent Feuillard, and Amandine Marrel. Numerical studies of the metamodel fitting and validation processes. *arXiv preprint arXiv :1001.1049*, 2010.

[68] Bertrand Iooss and Mathieu Ribatet. Global sensitivity analysis of computer models with functional inputs. *Reliability Engineering & System Safety*, 94(7) :1194–1204, 2009.

[69] Alexandre Janon, Thierry Klein, Agnes Lagnoux, Maëlle Nodet, and Clémentine Prieur. Asymptotic normality and efficiency of two sobol index estimators. *ESAIM : Probability and Statistics*, 18 :342–364, 2014.

[70] Alexandre Janon, Maëlle Nodet, and Clémentine Prieur. Confidence intervals for sensitivity indices using reduced-basis metamodels. *arXiv preprint arXiv :1102.4668*, 2011.

[71] Norman L Johnson. Bivariate distributions based on simple translation systems. *Biometrika*, pages 297–304, 1949.

[72] Ian T Jolliffe. Introduction to multiple time series analysis. *Technometrics*, 35(1) :88–89, 1993.

[73] Young-Ju Kim and Chong Gu. Smoothing spline gaussian regression : more scalable computation via efficient approximation. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 66(2) :337–356, 2004.

[74] A. Klimke and B. Wohlmuth. Algorithm 847 : Spinterp, piecewise multilinear hierarchical sparse grid interpolation in matlab. *ACM Transactions on Mathematical Software*, 31 :561–579, 2005.

[75] S Kucherenko, María Rodriguez-Fernandez, C Pantelides, and N Shah. Monte carlo evaluation of derivative-based global sensitivity measures. *Reliability Engineering & System Safety*, 94(7) :1135–1148, 2009.

[76] Sergei Kucherenko, Stefano Tarantola, and Paola Annoni. Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications*, 183(4) :937–946, 2012.

[77] Joseph C Lam, Kevin KW Wan, and Liu Yang. Sensitivity analysis and energy conservation measures implications. *Energy Conversion and Management*, 49(11) :3170–3177, 2008.

[78] Regis Lebrun and Anne Dutfoy. An innovating analysis of the nataf transformation from the copula viewpoint. *Probabilistic Engineering Mechanics*, 24(3) :312–320, 2009.

[79] Christiane Lemieux. *Monte carlo and quasi-monte carlo sampling*. Springer Science & Business Media, 2009.

[80] Kevin J Lomas and Herbert Eppel. Sensitivity analysis techniques for building thermal simulation programs. *Energy and buildings*, 19(1) :21–44, 1992.

[81] Henrik Madsen and Jan Holst. Estimation of continuous-time models for the heat dynamics of a building. *Energy and Buildings*, 22(1) :67–79, 1995.

[82] Andrzej Makagon. Stationary sequences associated with a periodically correlated sequence. *Probab. Math. Statist.*, 31(2) :263–283, 2011.

[83] T. A. Mara. Extension of the rbd-fast method to the computation of global sensitivity indices. *Reliability Engineering and System Safety*, 94 :1274–1281, 2009.

[84] Thierry A Mara and Stefano Tarantola. Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering & System Safety*, 107 :115–121, 2012.

[85] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[86] Houcem Eddine Mechri, Alfonso Capozzoli, and Vincenzo Corrado. Use of the anova approach for sensitive building energy design. *Applied Energy*, 87(10) :3073–3083, 2010.

[87] Max D Morris. Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2) :161–174, 1991.

[88] Roger B Nelsen. *An introduction to copulas*, volume 139. Springer Science & Business Media, 2013.

[89] J. E. Oakley and A. O'Hagan. Probabilistic sensitivity analysis of complex models : a bayesian approach. *J. Royal Stat. Soc. B*, 66 :751–769, 2004.

[90] Jouko Pakanen and Sami Karjalainen. Estimating static heat flows in buildings for energy allocation systems. *Energy and Buildings*, 38(9) :1044–1052, 2006.

[91] Giovanni Peccati. Hoeffding-anova decompositions for symmetric statistics of exchangeable observations. *Annals of probability*, pages 1796–1829, 2004.

[92] Aldomar Pedrini, Fernando Simon Westphal, and Roberto Lamberts. A methodology for building energy modelling and calibration in warm climates. *Building and Environment*, 37(8) :903–912, 2002.

[93] P. Pham Quang. *Modélisation magnéto-mécanique d'un nano commutateur. Optimisation sous contraintes de fiabilité par dérivation automatique des programmes en Java.* PhD thesis, 2011.

[94] E. Plischke. An effective algorithm for computing global sensitivity indices (easi). *Reliability Engineering and System Safety*, 95 :354–360, 2010.

[95] Carl Edward Rasmussen. Gaussian processes for machine learning. 2006.

[96] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, pages 832–837, 1956.

[97] Thierry Salomon, Renaud Mikolasek, and Bruno Peuportier. Outil de simulation thermique du bâtiment, comfie. *Journée thématique SFT-IBPSA mars*, 2005.

[98] A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computational Physics Communications*, 145 :280–297, 2002.

[99] A. Saltelli, K. Chan, and E. M. Scott. *Sensitivity Analysis*. Wiley, 1 edition, December 2008.

[100] A. Saltelli, S. Tarantola, and K. Chan. A quantitative model independent method for global sensitivity analysis of model output. *Technometrics*, 41 :39–56, 1999.

[101] A. Saltelli, Stefano Tarantola, Francesca Campolongo, and Marco Ratto. *Sensitivity Analysis in Practice : A Guide to Assessing Scientific Models*. John Wiley & Sons Ltd, February 2004.

[102] Andrea Saltelli, Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. *Global sensitivity analysis : the primer*. John Wiley & Sons, 2008.

[103] F. E. Satterthwaite. Random balance experimentation. *Technometrics*, 1 :111–137, 1959.

[104] J.H. Schaibly and K.E. Shuler. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. ii. applications. *J. Chemical Physics*, 59 :3879–3888, 1973.

[105] Tapio Schneider and Arnold Neumaier. Algorithm 808 : Arfit—a matlab package for the estimation of parameters and eigenmodes of multivariate autoregressive models. *ACM Transactions on Mathematical Software (TOMS)*, 27(1) :58–65, 2001.

[106] M Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Université Paris 8, 1959.

[107] S.A. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. *Doklady Akademii Nauk SSSR*, 4 :240–243, 1963.

[108] I. M. Sobol'. Sensitivity estimates for nonlinear mathematical models. *Math. Mod. and Comput. Exp.*, 1 :407–414, 1993.

[109] Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3) :271–280, 2001.

[110] Ilya M Sobol and Sergei Kucherenko. Derivative based global sensitivity measures and their link with global sensitivity indices. *Mathematics and Computers in Simulation*, 79(10) :3009–3017, 2009.

[111] Il'ya Meerovich Sobol'. On sensitivity estimation for nonlinear mathematical models. *Matematicheskoe Modelirovanie*, 2(1) :112–118, 1990.

[112] Clara Spitz, Laurent Mora, Etienne Wurtz, and Arnaud Jay. Practical application of uncertainty analysis and sensitivity analysis on an experimental house. *Energy and Buildings*, 55 :459–470, 2012.

[113] Charles J Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, pages 118–171, 1994.

[114] B. Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability Engineering and System Safety*, 93 :964–979, 2008.

[115] Jian Sun and T Agami Reddy. Calibration of building energy simulation programs using the analytic optimization approach (rp-1051). *HVAC&R Research*, 12(1) :177–196, 2006.

[116] S. Tarantola, D. Gatelli, and T. A Mara. Random balance designs for the estimation of first-order sensitivity indices. *Reliability Engineering and System Safety*, 91 :717–727, 2006.

[117] M. A. Tatang, W. W. Pan, R. G. Prin, and G. J. McRae. An efficient method for parametric uncertainty analysis of numerical geophisical model. *J. Geophysics Research*, 102 :21925–21932, 1998.

[118] Wei Tian and Pieter De Wilde. Uncertainty and sensitivity analysis of building performance using probabilistic climate projections : A uk case study. *Automation in Construction*, 20(8) :1096–1109, 2011.

[119] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[120] Shengwei Wang and Xinhua Xu. Simplified building model for transient thermal performance estimation using ga-based parameter identification. *International Journal of Thermal Sciences*, 45(4) :419–432, 2006.

[121] GS Watson. Analysis of dispersion on a sphere. *Geophysical Supplements to the Monthly Notices of the Royal Astronomical Society*, 7(4) :153–159, 1956.

[122] Fernando Simon Westphal and Roberto Lamberts. Building simulation calibration using sensitivity analysis. In *Building Simulation*, volume 9, pages 1331–1338. Citeseer, 2005.